NBER WORKING PAPER SERIES

REGULATING THE DIRECTION OF INNOVATION

Joshua S. Gans

Working Paper 32741
http://www.nber.org/papers/w32741

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2024

Thanks to Daron Acemoglu and Kevin Bryan for helpful discussions. All errors remain my own. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

Regulating the Direction of Innovation
Joshua S. Gans
NBER Working Paper No. 32741
July 2024
JEL No. L51,O33

ABSTRACT

This paper examines the regulation of technological innovation direction under uncertainty about potential harms. We develop a model with two competing technological paths and analyze various regulatory interventions. Our findings show that market forces tend to inefficiently concentrate research on leading paths. We demonstrate that ex post regulatory instruments, particularly liability regimes, outperform ex ante restrictions in most scenarios. The optimal regulatory approach depends critically on the magnitude of potential harm relative to technological benefits. Our analysis reveals subtle insurance motives in resource allocation across research paths, challenging common intuitions about diversification. These insights have important implications for regulating emerging technologies like artificial intelligence, suggesting the need for flexible, adaptive regulatory frameworks.

Joshua S. Gans
Rotman School of Management
University of Toronto
105 St. George Street
Toronto ON M5S 3E6
and NBER
joshua.gans@rotman.utoronto.ca

1 Introduction

While technological progress and innovation are generally presumed by economists to be about significant improvements to productivity, quality of life and general well-being, there is hardly any technology that does not involve harm (Acemoglu and Johnson, 2023). While it is often the case that those harms are outweighed by the benefits, there are prominent examples where this was not the case. The chemist Thomas J. Midgley was responsible for the invention of freon for refrigeration and anti-knock petrol. The former led to the release of chlorofluorocarbons and an expansion in the Ozone hole over the South Pole. The latter led to lead pollution with health and other consequences. In each case, governments determined that the harms outweighed the benefits and regulated a shift to alternative, but at the time, more costly technologies such as hydrofluorocarbons and unleaded petrol. For each of these alternatives, government regulation both stimulated adoption and, before that, scientific research to improve the viability of those alternatives.

In other areas, switching once the harms were understood has proved more challenging. At the beginning of the 20th century, automobiles were powered by both electric and internal combustion engines, with the latter 'winning' out in terms of adoption. However, for many decades, it has been known that petrol-powered cars resulted in more pollution than electric vehicles (McLaughlin, 1954). Yet despite some regulatory interventions, it has only recently occurred that some significant degree of switching to the alternative has occurred. Similarly, at the beginning of its deployment following World War II, nuclear power generation could be undertaken by heavy water or light water reactors. As Cowan (1990) documents, heavy water reactors initially had lower ongoing costs than light water ones and were likely to involve better safety outcomes. However, light water reactors were chosen for development in nuclear-powered submarines, and this drove private companies to invest in that path as a means of providing civilian power generation (the exception being Canada, which still has a heavy water reactor). Light-water reactors advanced technologically while other designs lagged behind. This made those reactors ubiquitous until safety concerns led to the halting of new nuclear power generation altogether following the Three Mile Island disaster in the 1970s (Bryan, 2017). The implication here is that one technological path can establish leadership, making it costly to switch to others should harm become apparent.

At present, the potential harms involved in technological innovation are being actively discussed and legislated with regard to artificial intelligence (AI). In the past decade, due to advances in computational statistics using machine learning, there have been significant advances that have allowed machines to engage more accurately and over a wider range of prediction tasks that were previously possible (Agrawal et al., 2022). These advances have the potential impact on many tasks currently performed by human workers as well as potentially to involve unintended consequences that may generate harm to security and political processes (Russell, 2019). Learning from these historical examples, some have argued that governments need to pre-emptively regulate both the adoption and direction of research associated with AI (Acemoglu, 2021). It has been noted that as AI develops along one path, it may become difficult to scale back adoption ex post (Acemoglu and Lensman, 2024). Thus, there is a call for pre-emptive regulation. While some of this regulation involves increased ex-ante assessments of the dangers associated with AI adoption, many of the proposals have involved interventions in pushing AI development toward outcomes that are 'human-centric' and more controllable (Brynjolfsson, 2022).

Several theoretical papers in economics have established that market forces left alone may lead to distortions in the chosen direction of technological change away from paths that might be less harmful (Bryan and Lemus, 2017), or promote insufficient diversity in scientific effort across alternative paths (Acemoglu, 2011). Similarly, there are concerns more closely related to AI and automation that the market promotes less efficient and more harmful avenues for technological change (Acemoglu, 2023). While the mechanisms for such welfare sub-optimal technological change differ between these models, there is a common policy conclusion that the ex ante promotion of under-developed research paths would be welfare-improving.

This paper revisits these calls for ex ante regulation in light of the current debate regarding AI. The model presented here is inspired by that debate but is provided as an examination of a generic general-purpose technology. The modelling innovation relative to previous work explicitly takes into account that, at the outset, the potential harms regarding alternative technological paths or architectures are uncertain, and it is in the context of that uncertainty that any ex-ante regulatory intervention must be made. That said, there is the potential for learning about the harms, although it is argued that this learning occurs primarily on paths that achieve some degree of adoption so that harms, if any, can surface or be dismissed. Thus, a regulator considering intervention must take into account not only uncertainty along paths 'the market' may be pursuing but also regarding the alternative.

Two broad findings arise from this examination. First, while it is the case that diversifying scientific resources across research paths can advance lagging technologies and reduce the costs of interventions should harms emerge on one path, that is not the only means by which scientific research can provide insurance against net harmful technologies being adopted. The other avenue is to double down and provide more resources to the leading path so that it advances to a sufficient level that its adoption involves net benefits even if harms should become apparent. This pushes regulators away from diversifying scientific resources across paths.

Second, because of this, heavy-handed ex-ante interventions such as prohibiting adoption or scientific research along a path can involve high costs relative to interventions that allow for 'pricing in' of potential harm. Two such alternative interventions are considered: Pigouvian taxation and ex-post liability. It is shown that the latter, precisely because it not only internalises harm but the prospect of harm when there is uncertainty, pushes agents in the economy to make more socially optimal adoption and research choices.

The paper proceeds as follows. The next section sets up the baseline model involving two potential technological paths that have differential appeals to different sectors in the economy and for which the potential harms are unknown for each path. Section 3 then characterises the decentralised equilibrium outcomes where both harms and their potential are not taken into account by private decision-makers. Section 4 then considers the socially optimal allocation and notes that changing scientific resource allocations across paths to take into account potential harm involves subtle insurance motives that may not be equivalent to more diversity in scientific research direction. Section 5 then examines and compares the alternative regulatory instruments, both ex ante and ex post, that can be deployed by regulators. Section 6 considers extensions, while a final section concludes.

2 Model Set-Up

The economy produces a unique final good from a continuum of sectors $i \in [0, 1]$ in each period t according to the production function:

$$Y(t) = \int_0^1 Y_i(t)di$$

A representative consumer has linear preferences over this final good and discounts the future at a rate of $\rho > 0$.

Firms in each sector choose the technology to adopt; either new technology architecture, A or B.[1] $Q_j(t) > 0$ denotes the quality of technology $j \in \{A, B\}$ at time t. For notational convenience, $x_i(t) = 1$ if sector i adopts architecture A in period t and $x_i(t) = 0$ if it, instead, adopts B.[2]

The impact of an architecture j's quality on sectoral output is captured by a parameter

[1] Output for each sector in the absence of adopting the new technology is set at zero for simplicity. Acemoglu and Lensman (2024) consider a choice between an old and new technology where it is the new technology that carries potential external damage.

[2] The discussion here is of "sectoral" adoption of technology whereas, in the model below, it is individual firms within a sector choosing a technology. As each firm within a sector is identical, it will turn out, in equilibrium, that all firms in a sector make the same adoption choice in each period.

$\eta_{i,j} \geq 0$. For convenience, sectors are ordered so that $\eta_{i,A} = 1 - i$ and $\eta_{i,B} = i$; that is, the distribution of sector-specific productivities is uniform with half of the sectors having a higher quality-adjusted productivity in architecture A (B), $i \in [0, \frac{1}{2})$ ($i \in (\frac{1}{2}, 1]$).[3] Under these assumptions, the output of sector i can be written as:

$$Y_i(t) = x_i(t)(1-i)Q_A(t) + (1 - x_i(t))iQ_B(t)$$

Given the ordering assumptions on i, at any given time, a particular focus will be on $\hat{I}(t)$, which captures the productivity parameter for A such that all sectors, $i \leq \hat{I}(t)$ adopt A and all sectors $i > \hat{I}(t)$ adopt B, at time t.

2.1 Innovation

Improvements to the quality of any architecture j are enabled by innovation from a fixed pool of scientists, S; a continuum on the unit interval. Each scientist has one unit of effort that can be applied to one architecture or the other in each period. If scientists apply total innovative effort, $s_j(t)$, to j in period t, then with probability $h(s_j(t))$ this generates $Q_j(t+1) = Q_j(t) + \Delta$, for $\Delta > 0$, with $Q_j(t+1) = Q_j(t)$ otherwise. $h(.)$ is non-decreasing, concave, continuously differentiable and satisfies the Inada conditions, $\lim_{s_j \to 0} h'(s_j) = \infty$ and $\lim_{s_j \to 1} h'(s_j) = 0$.[4]

It is assumed that the innovation is rival but perfectly excludable for the incremental innovation; specifically, in selling an architecture with quality $Q_j(t)$, the previous quality level, $Q_j(t-1)$ is freely available to sectors. This is akin to a quality ladder model where each quality step is excludable for one period only.[5] Finally, it is assumed that, in the absence of any innovation, $Q_j = 0$.

2.2 Externalities

Note that A and B represent two distinct paths for a general-purpose technology that all sectors can use. In addition, both can potentially give rise to external effects. The size of the externality is assumed to depend on the number of sectors using a given architecture and their individual scale (using output as a proxy); that is, $E_{i,j}(t) = -\eta_{i,j}\delta_j$ (in units of the

[3] This distributional assumption simplifies notation but does not play an important role in the results below. The important characteristic is that sectors can be ordered according to the magnitude of their comparative advantages of A or B adoption., holding quality constant.

[4] Note that, due to the assumptions here, $s_j(t)$, corresponds to both total effort and the share of scientific effort devoted to j.

[5] See O'Donoghue et al. (1998) for a discussion of this assumption.

final good) where $\delta_j \geq 0$ is common across sectors.[6]

A key assumption is that the value of δ_j is ex ante uncertain. To keep the analysis simple, two assumptions are made. First, $\delta_j \in \{0, \delta\}$ for each j and $\mu \in (0,1)$ represents the (common) prior that a given architecture has $\delta_j = \delta$. Second, $\delta \geq \Delta$. This allows us to focus on the interesting case where if an architecture has only advanced modestly, it is optimal to abandon it.[7]

2.3 Time structure

Timing in the model consists of two time periods, $t \in \{1, 2\}$. At the beginning of each period, scientists engage in research to determine the quality of a given architecture in that period. Following the realisation of outcomes from that research in period $t = 1$, sectors choose whether to adopt one technology architecture or not. Each sector chooses to adopt one of the architectures by paying a sector-specific price, $p_{i,j}(t)$. That adoption generates a signal of whether the technology architecture results in harm through a data-generating process described below. Then, the process repeats in period 2, except that adoption resolves any remaining uncertainty regarding potential harm. The particular focus of this paper is on the policy-choices that are implemented at the beginning of period 2 following any signals generated at the end of period 1.

2.4 Pricing

Each architecture is assumed to be marketed by a scientist-owned monopolist in each period. Thus, if the number of scientists contributing to architecture A is $s_A(t)$ and $\hat{I}(t)$ sectors adopt A at t, those scientists share in $v_A(t)$ if an innovation is generated at the beginning of period t. Here,

$$v_A(t) = \int_0^{\hat{I}(t)} \hat{p}_{i,j}(t) di$$

and similarly,

$$v_B(t) = \int_{\hat{I}(t)}^1 \hat{p}_{i,j}(t) di$$

Given this, scientists can charge a sector-specific price, $p_{i,j}(t)$ where:

$$\hat{p}_{i,j}(t) = \eta_{i,j} Q_j(t) - \max\{\eta_{i,j} Q_j(t-1), \eta_{i,-j} Q_{-j}(t) - \hat{p}_{i,-j}(t)\}$$

[6] In a later section, the case where $E_{i,j}(t) = -\eta_{i,j}\delta_j Q_j(t)$ and damages scale with sectoral output is discussed.

[7] A case that is also the focus of Acemoglu and Lensman (2024). The impact of differing magnitudes of relative damage is discussed below.

All scientists contributing research effort to that architecture share equally in these commercial returns; thus, scientists contributing to A receive $\frac{v_A(t)}{s_A(t)}$ and those contributing to B receive $\frac{v_B(t)}{s_B(t)}$. This means that scientists with an innovation on a given architecture compete with other scientists with an innovation on the alternative architecture for adoption by a sector at any given time. As the competition is in price terms, then it is clear that, in equilibrium, at least, $\hat{p}_{i,A}(t)$ *or* $\hat{p}_{i,B}(t)$ will be zero for a given sector i (with both being zero if their characteristics are identical).

3 Decentralised Equilibrium

Without regulation, there is no incentive for scientists and firms to consider the external effects of technology architectures should they exist. Let $\{q_A(t), q_B(t)\}$ be the state of each architecture where $Q_A(t) = q_A(t)\Delta$ and $Q_B(t) = q_B(t)\Delta$ and each $q_j(t) \in \{0, 1, 2\}$. Note that because there are no constraints on firms switching architectures in each period, in equilibrium, the threshold, $\hat{I}(q_A(t), q_B(t))$, defining the sectors adopting A (viz B) is defined by:

$$(1 - \hat{I}(q_A(t), q_B(t))Q_A(t) = \hat{I}(q_A(t), q_B(t))Q_B(t) \implies \hat{I}(q_A(t), q_B(t)) = \frac{Q_A(t)}{Q_A(t) + Q_B(t)}$$

Another threshold of interest concerns the outside options of firms in each sector when considering adopting an architecture. For instance, a firm in sector i who does not purchase an improvement to architecture A will purchase the alternative architecture only if $(1 - \eta_{i,A})Q_B(t)$ is greater than $\eta_{i,A}Q_A(t-1)$. Thus, we can define $\hat{I}_A(q_A(t), q_B(t))$ as:

$$\hat{I}_A(q_A(t), q_B(t))Q_A(t-1) = (1 - \hat{I}_A(q_A(t), q_B(t)))Q_B(t)$$
$$\implies \hat{I}_A(q_A(t), q_B(t)) = \frac{Q_A(t-1)}{Q_A(t-1) + Q_B(t)}$$

Note that $\hat{I}_A(q_A(t), q_B(t)) \leq \hat{I}(q_A(t), q_B(t))$ and, thus, sectors $i \leq \hat{I}_A(q_A(t), q_B(t))$ have an outside option from not purchasing the improved quality $Q_A(t)$ as continuing to use the previous quality (which is freely available) while those $i \in [\hat{I}_A(q_A(t), q_B(t)), \hat{I}(q_A(t), q_B(t))]$ have an outside option of purchasing the current best version of architecture B.

3.1 Scientist Allocation at $t = 1$

As scientist and firm choices in period $t = 1$ do not constrain their choices in the next period $t = 2$, they make choices to optimise their current payoffs. Consider the decision of firms to adopt architecture A. Firms in sector i, if offered a technology with quality $Q_A(1)$, will only purchase it only if $p_{i,A} \leq \eta_{i,A} Q_A(1)$. Moreover, if the current best available quality for B is $Q_B(1)$, they will purchase A only if $p_{i,A} \leq \eta_{i,A} Q_A(1) - \eta_{i,B} Q_B(1)$.

Scientists only generate a return if they produce an innovation in the current period. The magnitude of that return will depend upon whether scientists working on the alternative architecture have produced an innovation or not. Suppose that innovations arise on both paths so that $Q_A(1) = Q_B(1) = \Delta$. In this case, as an advance is required for production, the only constraint on pricing technologies based on one architecture is that based on the other. Thus, a sector, i, will adopt architecture A if $\eta_{i,A}\Delta - \hat{p}_{i,A} \geq \eta_{i,B}\Delta - \hat{p}_{i,B}$. Let $\hat{I}(1,1) = \{i|\eta_{i,A} = \eta_{i,B}\}$. Note that for all sectors, $i \leq \hat{I}(1,1)$, $\hat{p}_{i,B} = 0$. Thus, the maximum value of $\hat{p}_{i,A}$ is $(\eta_{i,A} - \eta_{i,B})\Delta$. A similar calculation shows that for $i > \hat{I}(1,1)$, $\hat{p}_{i,B} = (\eta_{i,B} - \eta_{i,A})\Delta$. Note that because $Q_A(1) = Q_B(1)$, $\hat{I}(1,1) = \frac{1}{2}$. Thus, the total A-scientist returns are:

$$v_A(1,1) = \int_0^{\frac{1}{2}} (\eta_{i,A} - \eta_{i,B})\Delta \, di$$

Recalling that $\eta_{i,A} = 1 - i$ and $\eta_{i,B} = i$, it can be seen that:

$$v_A(1,1) = \int_0^{\frac{1}{2}} (1 - 2i)\Delta \, di = \frac{\Delta}{4}$$

$v_B(1,1)$ has the same value. Suppose that the outcomes of $t = 1$ research are $Q_A(1) = \Delta$ while B has not advanced. Using a similar calculation for the case where both paths have advanced and noting that $\hat{I}(1,0) = 1$, we can derive:

$$v_A(1,0) = \int_0^{1} (1 - i)\Delta \, di = \frac{\Delta}{2}$$

Clearly, without the competitive pressure from architecture B, $v_A(1,0)$ exceeds $v_A(1,1)$. In this case, $v_B(1,0) = 0$.

To determine the equilibrium allocation of scientists to each architecture, note that the expected returns to each research path are:

$$V_A(0,0) = \frac{h(s_A(1))}{s_A(1)} \Big(h(s_B(1)) v_A(1,1) + (1 - h(s_B(1))) v_A(1,0) \Big)$$

$$V_B(0,0) = \frac{h(s_B(1))}{s_B(1)} \Big(h(s_A(1)) v_B(1,1) + (1 - h(s_A(1))) v_B(1,0) \Big)$$

Each scientist will choose a path that earns them the highest expected return. Let $s_A = s$ and $s_B = 1 - s$. Then, as $v_A(.) = v_B(.)$, the only point where $V_A(0,0) = V_B(0,0)$ is where

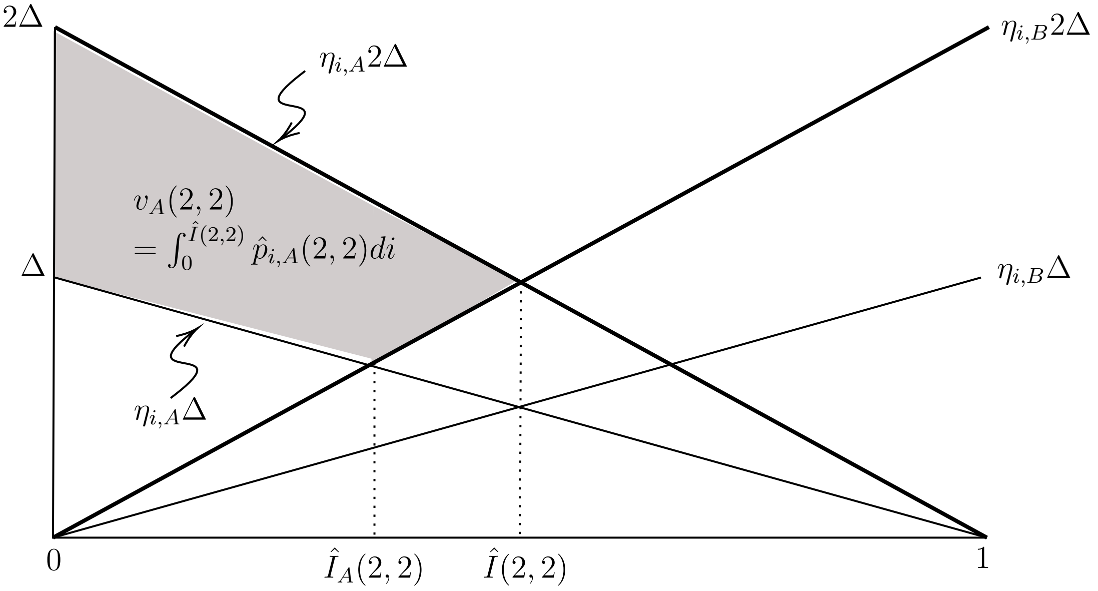

Figure 1: **Scientist Returns to Architecture A for $Q_A(2) = Q_B(2) = 2\Delta$**

$s = \frac{1}{2}$. Thus, scientists will allocate themselves in equal numbers to each path in equilibrium, i.e., $\hat{s}(1) = \frac{1}{2}$, as otherwise scientists would have an incentive to switch to the path offering the highest average return.

3.2 Scientist Allocation at $t = 2$

At $t = 2$, there are four possible outcomes for the state of the architecture at the beginning of the period: $\{q_A, q_B\}$ could be $\{1, 1\}$, $\{0, 0\}$, $\{1, 0\}$ or $\{0, 1\}$. If a previous innovation occurred during $t = 1$, it is now in the public domain, and any sector can utilize it at no cost.

This has an impact on the pricing that can be achieved should there be innovations in $t = 2$; in particular, sectors with a strong preference for a particular architecture may prefer to continue to use the previous generation of technology rather than the alternative architecture, even if it has advanced. To see this, suppose that $Q_A(2) = Q_B(2) = 2\Delta$ (i.e., both architectures have advanced from a starting point where $(q_A, q_B) = (1, 1)$). In a price-setting game, all sectors where $i \leq \hat{I}(2, 2)$ adopt $Q_A(2)$, with $\hat{p}_{i,A} = \eta_{i,A}\Delta$ for $i \leq \hat{I}_A(2, 2)$ and $\hat{p}_{i,A} = (\eta_{i,A} - \eta_{i,B})\Delta$ for $i \in [\hat{I}_A(2, 2), \hat{I}(2, 2)]$. This outcome is depicted in Figure 1. Note that because $Q_A(2) = Q_B(2)$, $\hat{I}(2, 2) = \frac{1}{2}$. Thus, total A-scientist returns are:

$$v_A(2, 2) = \int_0^{\hat{I}_A(2,2)} \eta_{i,A}\Delta \, di + \int_{\hat{I}_A(2,2)}^{\frac{1}{2}} (\eta_{i,A} - \eta_{i,B})2\Delta \, di$$

Recalling that $\eta_{i,A} = 1 - i$ and $\eta_{i,B} = i$, note that, in this case, $\hat{I}_A(2,2) = \{i|(1-i)\Delta = i2\Delta\} = \frac{1}{3}$ which leads to:

$$v_A(2,2) = \int_0^{\frac{1}{3}} (1-i)\Delta\, di + \int_{\frac{1}{3}}^{\frac{1}{2}} (1-2i)2\Delta\, di = \frac{\Delta}{3}$$

Again, considering a starting point where $\{q_A, q_B\} = \{1,1\}$, if only one architecture, say A, has an innovation, then it is easy to calculate that $\hat{I} = \frac{1}{2}$ and $\hat{I}_A(2,1) = \frac{2}{3}$. Therefore,

$$v_A(2,1) = \int_0^{\frac{1}{2}} (1-i)\Delta\, di + \int_{\frac{1}{2}}^{\frac{2}{3}} ((1-i)2\Delta - i\Delta)\, di = \frac{5}{12}\Delta$$

(Note that we state $v_A(q_A, q_B)$ and $v_A(q_A, q_B)$ both as functions of the state $\{q_A, q_B\}$.) To determine the equilibrium allocation of scientists to each architecture, note that, at the beginning of $t = 1$ when $\{q_A, q_B\} = \{1,1\}$, the expected returns to each research path are:

$$V_A(1,1) = \frac{h(s_A)}{s_A}\Big(h(s_B)v_A(2,2) + (1 - h(s_B))v_A(2,1)\Big) = \frac{h(s)}{s}\tfrac{1}{2}\Big(1 - \tfrac{1}{3}h(1-s)\Big)\Delta$$

$$V_B(1,1) = \frac{h(s_B)}{s_B}\Big(h(s_A)v_B(2,2) + (1 - h(s_A))v_B(1,1)\Big) = \frac{h(1-s)}{1-s}\tfrac{1}{2}\Big(1 - \tfrac{1}{3}h(s)\Big)\Delta$$

Given the symmetry involved, the equilibrium allocation involves $\hat{s}(2) = \frac{1}{2}$.

Note that if $\{q_A, q_B\} = \{0,0\}$, then the possible outcomes at $t = 2$ are the same as those at $t = 1$. However, if, say, $\{q_A, q_B\} = \{1,0\}$, if both paths advance at $t = 2$, then, $v_A(2,1) = \frac{5}{12}\Delta$ as derived above while

$$v_B(2,1) = \int_{\frac{2}{3}}^{1} (i\Delta - (1-i)2\Delta)\, di = \frac{\Delta}{6}$$

If only one path advances at $t = 2$, then $v_B(1,1) = \frac{\Delta}{4}$ as derived above while

$$v_A(2,0) = \int_0^1 (1-i)\Delta\, di = \frac{\Delta}{2}$$

Given this, the expected returns to each research path are:

$$V_A(1,0) = \frac{h(s_A)}{s_A}\Big(h(s_B)v_A(2,1) + (1 - h(s_B))v_A(2,0)\Big) = \frac{h(s)}{s}\tfrac{1}{2}\Big(1 - \tfrac{1}{6}h(1-s)\Big)\Delta$$

$$V_B(1,0) = \frac{h(s_B)}{s_B}\Big(h(s_A)v_B(2,1) + (1 - h(s_A))v_B(1,1)\Big) = \frac{h(1-s)}{1-s}\tfrac{1}{4}\Big(1 - \tfrac{1}{3}h(s)\Big)\Delta$$

As $V_A(1,0) > V_B(1,0)$ for $s = \frac{1}{2}$, it is clear that the equilibrium involves $\hat{s}(2) > \frac{1}{2}$. That is, more scientists are allocated to the leading architecture than the lagging one. Intuitively, the more advanced path can potentially earn a higher return than the other path regardless of whether the other path advances or not, whereas the less advanced path can, at best, catch up commercially. Hence, more scientists will choose to research in the more advanced path. Interestingly, this outcome arises even though there are no constraints on sectors switching their adopted architecture from their previously chosen architecture nor any constraints on scientists in switching research paths.[8]

4 Socially Optimal Allocations

Two potential sources of social inefficiencies arise in the equilibrium without regulation. First, firms and scientists do not consider externalities (if any) associated with any architecture they may research or adopt. Second, firms and scientists are not long-lived; specifically, their decisions at $t = 1$ do not take into account how this impacts the diversity of options available for firms to adopt at $t = 2$. Here we consider the socially optimal allocation of scientists to alternative research paths.

The social planner chooses $\{s_A(t), s_B(t), \{x_i(t)\}_{i\in[0,1]}\}_{t=1,2}$ to maximise:

$$\sum_{t=1}^{2} \rho^{t-1}\mathbb{E}[Y(t) - E(t)]$$

Note, however, that since the level of the externality is unknown at $t = 1$, allocations at that time will be based on its expected value, while those at $t = 2$ may take into account the realised value of the externality.

4.1 Adoption at $t = 2$

Consider the social planner's choice of which sectors should adopt which architecture following a realisation of the magnitude of any externalities. Suppose that the realised values of the relevant parameters are $\{\delta_A, \delta_B\}$; recalling that these are common across sectors. Let I denote the threshold whereby sectors, $i \leq I$ adopt A and the remainder adopt B. In this

[8] In contrast to other models where innovation takes place on technologies at different steps on a quality ladder, no scientist or firm owns or is tied to a particular architecture. Moreover, past innovations are freely available.

case, the social planner chooses I to maximise:

$$\int_0^I \eta_{i,A}(Q_A(2) - \delta_A)di + \int_I^1 \eta_{i,B}(Q_B(2) - \delta_B)di$$

The optimum satisfies:

$$I^* = \frac{Q_A(2) - \delta_A}{Q_A(2) - \delta_A + Q_B(2) - \delta_B}$$

To focus on a case of interest, suppose that the maximum realised externalities are such that $\delta_A, \delta_B \leq 2\Delta$.[9] Thus, at the lowest quality level for an architecture, it is still optimal for the sector that most benefits from that architecture to adopt that technology. Thus, regardless of quality, I^* has the interior solution given above.

Note that when $\delta_A = \delta_B$ (including the case where both are 0), the optimal adoption level, $I^* = \hat{I}$, the equilibrium adoption level. Thus, it is only when the architectures have different realised externalities that it is optimal to deviate from the equilibrium adoption level and favour adoption by sectors of the architecture with the lowest externality.

At I^*, the social welfare realised at $t = 2$ is:

$$v_S(q_A, q_B, \delta_A, \delta_B) = \frac{1}{2}\left(\frac{(q_A\Delta - \delta_A)^2 + (q_B\Delta - \delta_B)^2 + (q_A\Delta - \delta_A)(q_B\Delta - \delta_B)}{q_A\Delta - \delta_A + q_B\Delta - \delta_B}\right)$$

where we recall that $Q_j(2) = q_j\Delta$.

4.2 Scientist Allocation at $t = 2$

The socially optimal allocation of scientists will be impacted by two factors that differ from a decentralised allocation. One is, of course, knowledge of the extent of any externalities that, as noted above, impacts on the socially optimal adoption of technology. The other is that, in a decentralised allocation, scientists choose their research paths based on relative average expected returns, whereas, as will be demonstrated, the socially optimal allocation depends on relative marginal expected returns. In order to build intuition, it is instructive first to consider the case where there are no externalities (i.e., either μ or δ are zero) before moving to move beyond this case.

[9] Below alternative assumptions are considered, noting that the core assumption here captures the most cases of interest.

4.2.1 No externalities

In the absence of externalities, based on the current technology state, $\{q_A, q_B\}$ at the beginning of $t = 2$, the social planner chooses s to maximise:

$$\begin{aligned} V^*(q_A, q_B, 0, 0) \equiv \mathbb{E}[v_S(q_A, q_B, 0, 0)] = {} & h(s)h(1-s)v_S(q_A+1, q_B+1, 0, 0) \\ & + h(s)(1-h(1-s))v_S(q_A+1, q_B, 0, 0) \\ & + (1-h(s))h(1-s)v_S(q_A, q_B+1, 0, 0) \\ & + (1-h(1-s))(1-h(s))v_S(q_A, q_B, 0, 0) \end{aligned}$$

The assumptions on $h(.)$ ensure an interior equilibrium. Thus, s^* satisfies the first-order condition:

$$\frac{h'(s^*)}{h'(1-s^*)} = \frac{\Omega(q_A, q_B, 0, 0)h(s^*) - (v_S(q_A, q_B+1, 0, 0) - v_S(q_A, q_B, 0, 0))}{\Omega(q_A, q_B, 0, 0)h(1-s^*) - (v_S(q_A+1, q_B, 0, 0) - v_S(q_A, q_B, 0, 0))}$$

where $\Omega(q_A, q_B, 0, 0)$ is the degree of substitutability between A and B, i.e.:

$$\Omega(q_A, q_B, 0, 0) \equiv v_S(q_A, q_B+1, 0, 0) + v_S(q_A+1, q_B, 0, 0) - v_S(q_A+1, q_B+1, 0, 0) - v_S(q_A, q_B, 0, 0)$$

From this equation, it is clear that if $q_A = q_B$, then $s^* = \frac{1}{2}$ and scientists are equally allocated across research paths just as in the decentralised equilibrium.

The case of interest, therefore, is when $q_A \neq q_B$. To explore this, suppose that $q_A > q_B$ so that at the beginning of $t = 2$, A is the leading architecture, and B is the lagging architecture. While pursuing research on each path involves an increment to quality of Δ if either architecture is improved, the intermediate sectors, which are more indifferent between the two architectures, benefit as they do if both are improved. Thus, from a social welfare perspective, improvements are substitutes with $\Omega(1, 0) = \frac{1}{12}\Delta$. Importantly, $v_S(2, 0, 0, 0) = \Delta > v_S(1, 1, 0, 0) = \frac{3}{4}\Delta$ implying that more scientific resources are allocated to path A, the leading architecture, than B, the lagging architecture. Thus, we have:

$$\frac{h'(s^*(1,0))}{h'(1-s^*(1,0))} = \frac{1 - \frac{1}{3}h(s^*(1,0))}{2 - \frac{1}{3}h(1-s^*(1,0))}$$

It is easy to see that at $s = \frac{1}{2}$, the left-hand-side is less than 1. As $h(.)$ is concave this implies that $s^*(1, 0) > \frac{1}{2}$. Improvements in the leading architecture have a higher marginal return than for the lagging architecture as they have the potential to be spread across a broader measure of sectors.

This outcome can be compared with the decentralised equilibrium scientist allocation.

Note that, at that equilibrium,

$$V_A(1,0) = V_B(1,0) \implies \frac{h(\hat{s}(1,0))/\hat{s}(1,0)}{h(1-\hat{s}(1,0))/(1-\hat{s}(1,0))} = \frac{1-\frac{1}{3}h(\hat{s}(1,0))}{2-\frac{1}{3}h(1-\hat{s}(1,0))}$$

The RHS of this equation has the same structure as that for the earlier condition for the social optimum. The difference is in the LHS which for the social optimum is the ratio of marginal probabilities, whereas for the decentralised allocation, it is the ratio of average probabilities of success on each respect research path.

The following lemma characterises the relationship between the socially optimal and decentralised allocations when there are no externalities.

Lemma 1 *Suppose that $\frac{h'(s)}{h(s)/s}$ is decreasing in s. Then (i) if $s^* = \frac{1}{2}$, $\hat{s} = s^*$; and (ii) if $s^* > (<)\frac{1}{2}$, $\hat{s} > (<)s^*$.*

The proof is in the appendix. Lemma 1 demonstrates that the decentralised allocation typically results in a more concentrated allocation of scientists to one research path, which is the same one that attracts a higher allocation of scientists in the socially optimal allocation. Thus, the decentralised allocation is *amplified* relative to the socially optimal allocation. These outcomes are depicted graphically in Figure 2. The reason is that when scientists choose between alternative research paths, they are considering the relative *average* probabilities of success on that path, and so neglect the impact that their own choice has on the relative *marginal* probabilities of success. In particular, the probability of success for each path depends on each scientist's choice, but when a scientist switches from the lagging to the leading path, the marginal negative impact on the less advanced path is higher than the marginal positive impact on the probability that the more advanced path succeeds in advancing further. The social planner takes these impacts into account, while scientists only consider the latter positive impact on the path they pursue.

The result here is that there is socially too high a degree of concentration on the leading research path that the lagging in a decentralised equilibrium arises from a distinct rationale than similar results in the literature. For instance, Acemoglu (2011) provides a model where a leading path has an advantage in that products along this path can be commercialised immediately while those on a lagging path may have to wait until some event, such as changing tastes, makes them commercially viable. The market then underprovides diversity due to the asymmetric nature of private appropriation along the two competing paths. Bryan and Lemus (2017) show that racing distortions, the value of being first to advance a technological path, confers a negative externality on research incentives on the other path that is not taken into account by scientists causing them to allocate too many resources to "quick wins." Sim-

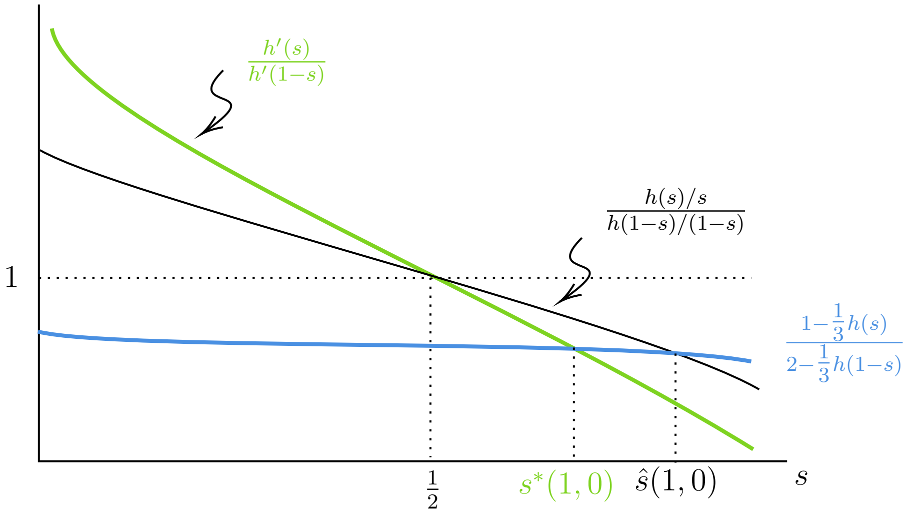

Figure 2: Socially Optimal versus Decentralised Scientist Allocations for $\{q_A, q_B\} = \{1, 0\}$

ilarly, scientists do not necessarily place sufficient value on the depth of research paths – in terms of who many future innovations they might yield – and so may inefficiently devote too many scientific resources to paths where innovation opportunities are front-loaded in time. While Bryan and Lemus (2017)'s results do not specifically address issues of the optimal level of diversity in research, they do identify some key distortions that arise. By contrast to both of these papers, here the inefficient concentration of research following asymmetric advances in period 1 arises because scientists fail to consider the negative externality imposed on other scientists if they switch from the lagging to the leading path and, thus, represents a distinct but complementary externality to those already identified in the literature. It is easy to imagine that a different model specification could emphasise a different externality in scientist allocation without any change to the broad conclusions reached below.

It is useful to highlight some functional forms for $h(.)$ that violate the assumed conditions for Lemma 1 to examine their role. First, if $h(s) = s^a$ for $a \in (0, 1)$, $\frac{h'(s)}{h(s)/s} = a$ for all s. In this case, the ratio of the marginal probabilities for each path equals the ratio of average probabilities, and so $s^* = \hat{s}$ even where $q_A \neq q_B$. Thus, the condition in the lemma ensures the potential for a divergence between socially optimal and decentralised allocations. Second, if $h(s) = \log(1 + s)$, then $\frac{h'(s)}{h(s)/s}$ is decreasing in s but the Inada conditions do not hold; namely, $h'(0) = 1$ and $h'(1) = 2$. While this does not alter the result that $s^* = \hat{s} = \frac{1}{2}$ when $q_A = q_B$, when, say, $\{q_A, q_B\} = \{1, 0\}$, a corner solution arises for both the decentralised equilibrium and social optimum with $s^* = \hat{s} = 1$. Thus, the Inada conditions are necessary for establishing an interior allocation.

4.2.2 Externalities

When technology adoption potentially results in direct harm, as discussed above, the socially optimal adoption choices of the technology can differ from the decentralised equilibrium adoption. Specifically, if it is known that $\delta_j = \delta$, then the social planner will not want to adopt j in any sector unless it has advanced to $Q_j = 2\Delta$.

However, at the beginning of $t = 2$, the social planner is uncertain regarding the nature of externalities but knows the current technology state, $\{q_A, q_B\}$ and may, relying on signals yet to be specified, have updated probabilities regarding the level of any externalities. Thus, when choosing the allocation of scientists to each path, it chooses s to maximise:

$$\begin{aligned}\mathbb{E}[V^*(q_A, q_B, \delta_A, \delta_B)] =& h(s)h(1-s)\mathbb{E}[v_S(q_A+1, q_B+1, \delta_A, \delta_B)]\\ &+ h(s)(1-h(1-s))\mathbb{E}[v_S(q_A+1, q_B, \delta_A, \delta_B)]\\ &+ (1-h(s))h(1-s)\mathbb{E}[v_S(q_A, q_B+1, \delta_A, \delta_B)]\\ &+ (1-h(1-s))(1-h(s))\mathbb{E}[v_S(q_A, q_B, \delta_A, \delta_B)]\end{aligned}$$

where the expectations here are with respect to $\{\delta_A, \delta_B\}$. The corresponding first-order condition for s^* is:

$$\frac{h'(s^*)}{h'(1-s^*)} = \frac{\Omega(q_A, q_B)h(s^*) - (\mathbb{E}[v_S(q_A, q_B+1, \delta_A, \delta_B)] - \mathbb{E}[v_S(q_A, q_B, \delta_A, \delta_B)])}{\Omega(q_A, q_B)h(1-s^*) - (\mathbb{E}[v_S(q_A+1, q_B), \delta_A, \delta_B)] - \mathbb{E}[v_S(q_A, q_B, \delta_A, \delta_B)])}$$

Here is the degree of substitutability between A and B, $\Omega(q_A, q_B)$, is now stated taking into account the planner's uncertainty:

$$\begin{aligned}\Omega(q_A, q_B) \equiv& \mathbb{E}[v_S(q_A, q_B+1, \delta_A, \delta_B)] + \mathbb{E}[v_S(q_A+1, q_B, \delta_A, \delta_B)]\\ &- \mathbb{E}[v_S(q_A+1, q_B+1, \delta_A, \delta_B)] - \mathbb{E}[v_S(q_A, q_B, \delta_A, \delta_B)]\end{aligned}$$

It is clear that if $q_A = q_B$ and $\mathbb{E}[\delta_A] = \mathbb{E}[\delta_B]$, then $s^* = \frac{1}{2}$ and scientists are equally allocated across research paths just as in the decentralised equilibrium.

What if there are some asymmetries between the two paths? Consider, first, differences in progress along each path; i.e., the case where $\{q_A, q_B\} = \{1, 0\}$. When, at that stage, there is still uncertainty regarding whether externalities might arise on each path, then a new social motive guiding the allocation of scientists emerges: *insurance*. However, the insurance motive is subtler than common intuition might have suggested. Common intuition on insurance is that diversification is valuable because it provides additional options should harm emerge on a research path.

To see this, suppose that for each path, at $t = 2$, the belief that a path will be harmful remains symmetric and equal to μ. (Below, it will be argued that this is unlikely as there will be some learning and updating, especially on the A path, but it is a useful starting point). While $\Omega(1,0)$ remains positive when there are externalities,[10] $\Omega(1,0)$ is decreasing in δ, implying that A and B become less substitutable as the size of the externality rises. Note, however, that $\mathbb{E}[v_S(2,0,\delta_A,\delta_B)] = \Delta - \mu\frac{1}{2}\delta > \mathbb{E}[v_S(1,1,\delta_A,\delta_B)] = \frac{1}{4}(1-\mu)(\mu+3)\Delta$ with the difference between them decreasing in δ. That is, as the size of the externality increases, it is optimal to reduce the share of scientists allocated to the leading path (A). The following proposition summarises the comparative statics.

Proposition 1 *Suppose that $q_A > q_B$ and $\mathbb{E}[\delta_A] = \mathbb{E}[\delta_B] = \mu$, then $s^*(2) > \frac{1}{2}$. $s^*(2)$ is decreasing in δ, increasing in Δ, and, for μ sufficiently high, increasing in μ.*

The insurance motive pushes scientific resources towards the lagging technology when the potential loss from externalities becomes high as the costs can potentially be spread across more sectors ex post. The reverse intuition holds for a higher Δ as this improves the return to advancing the leading architecture relative to having the lagging, substitute architecture catch up. As for the probability of an externality arising, μ, when μ is very high, the likelihood of an externality is high, in which case the lagging architecture serves less of an insurance role. Specifically, a poor outcome is highly likely on both architectures, but the leading architecture, should it advance, can weather the externality. By contrast, when μ is low, an increase in μ, raises the insurance value of allocating scientific resources towards developing the lagging technology.

When there are externalities, how does the socially optimal allocation compare with the decentralised outcome? Recall that, in the absence of direct externalities, the socially optimal allocation involved a less concentrated allocation than the decentralised outcome. When there are possible direct externalities, a simple intuition might suggest that the social planner has a stronger incentive to prefer a more balanced allocation of scientists between the two research paths as this mitigates the risk of advancing a path that also involves higher realised external harm. However, this simple intuition turns out to be incorrect.

To see this, consider the case where it is maximally uncertain as to whether externalities will arise or not; i.e., where $\mu = \frac{1}{2}$. For this case, the following proposition can be proved.

Proposition 2 *Suppose that $\mu = \frac{1}{2}$. If $\delta \geq \frac{3}{2}\Delta$, then $s^* \leq \hat{s}$. If $\delta < \frac{3}{2}\Delta$, it is possible that $s^* > \hat{s}$.*

The proof is straightforward and omitted. The proposition shows that a sufficient condition for the simple, intuitive result to arise involves the potential harm being sufficiently high (i.e.,

[10] Specifically, $\Omega(1,0) = \frac{\Delta(1-\mu)(3(\Delta+3\Delta\mu)-\delta(1+5\mu))}{12(3\Delta-\delta)}$.

$\delta \geq \frac{3}{2}\Delta$). However, if the potential harm is relatively low (i.e., $\delta < \frac{3}{2}\Delta$), this is a necessary condition for the socially optimal allocation to be more concentrated than the decentralised allocation. Intuitively, the best way to insure against the consequences of harm is to advance the leading architecture further so that it can 'pay' for the harm by being adopted. This is not an option for the lagging architecture. Therefore, it is an insurance motive but the way in which insurance is provided is through an acceleration in research along one path rather than diversification across paths.

The following result provides a case where the conditions in Proposition 2 can be necessary and sufficient.

Corollary 2 *Suppose that $\mu = \frac{1}{2}$ and $h(s) = s^a$ (with $a < 1$). $s^* \leq \hat{s}$ if and only if $\delta \geq \frac{3}{2}\Delta$.*

What this shows is for a case where the distortion from the decentralised allocation of scientists is not present, the impact of direct externalities is clearer.

In effect, the insurance motive is subtle because insurance can be obtained both by diversifying across paths *and* by investing more in the leading path, which can still be valuable even if harm arises. Indeed, we can go further and show that diversified investment is still desirable when it is known that one path involves harm, but the other does not. For instance, suppose that $q_A = q_B = 1$ but that $\delta_A = 1$ and $\delta_B = 0$. A simple intuition would suggest that, in this case, s^* should equal 0 with all resources devoted to advancing B (which, if this advance occurs, generates a surplus of $v_S(0, 2, \delta, 0) = \Delta$. However, it is possible that B does not advance. In this case, only a B technology product is available (the A line not advancing and, therefore, being inefficient to adopt at all), and so total surplus is $v_S(1, 1, \delta, 0) = \frac{\Delta}{2}$. By contrast, if some scientific resources are devoted to the B technology, then there is some probability that both technologies advance generating surplus of $v_S(2, 2, \delta, 0) = \frac{1}{2}(2\Delta + \frac{(2\Delta-\delta)^2}{4\Delta-\delta}$ or only A advances yielding $v_S(2, 1, \delta, 0) = \frac{1}{2}(2\Delta + \frac{(2\Delta-\delta)^2}{3\Delta-\delta}$ both of which exceed $\frac{\Delta}{2}$. Thus, as $h(.)$ is concave, then $s^* > 0$ even when harms are known and asymmetric. Intuitively, the A technology has a higher value for some sectors, which is realised, even with harm, if that technology advances.

What this section demonstrates is that from the perspective of allocating scientific resources, there are no clear courses of action when uncertainty is resolved and even more subtle interactions when uncertainty remains. This means that there are challenges and trade-offs involved between alternative instruments for intervention; something which will be addressed next.

5 Regulatory Interventions

There is only an incentive to regulate the direction of technological change if the planner receives some signal of the potential harm that might arise from one path. Absent this, in this environment, both paths look equally attractive, and so there is no basis on which to favour one of the other. Nonetheless, there is still potential for both ex ante and ex post intervention if a signal of harm is received. In what follows, the policy options based on such a signal are analysed.

Four policy options are available: (i) ban the adoption of an architecture, (ii) ban research advancing an architecture, (iii) impose a tax on the adoption of an architecture of $E_{i,j}$ and (iv) subjecting adopters to ex post liability of $E_{i,j}$ if harm occurs. The first three of these are forms of ex ante regulation as they regulate prior to the realisation of research outcomes in period 2. The final outcome is an ex post regulation because regulation only proceeds after all activity has occurred. This section will evaluate and compare each of these options contingent on the receipt of various policy signals outlined next.

5.1 Policy Signals

Before analysing the impact of the policy options, it is important to more precisely specify the data-generating process for policy signals that the policy-maker might receive at the end of period 1 that would trigger various policy outcomes in period 2. The core assumption made here is that actual adoption of a technology architecture is required to receive a signal about that architecture's potential harmfulness.[11]

Suppose that in period 1, the scale of use of technology A is I_1 and that of B is $1 - I_1$. Recall that the prior probabilities that harm arises are the same across architectures and are μ for each. Moreover, these probabilities are independent. Given I_1, if there is no signal of harm, that probability is updated according to Bayes' rule as given by the following formula:

$$\tilde{\mu}_A = \frac{\mu(1 - I_1)}{\mu(1 - I_1) + (1 - \mu)}$$

$$\tilde{\mu}_B = \frac{\mu I_1}{\mu I_1 + (1 - \mu)}$$

where $\tilde{\mu}_j$ is the (posterior) probability that damage could arise from the use of j in $t = 2$ given that it has not arisen prior to that point in $t = 1$. The learning process assumed here is that unless damage is observed during a period, the likelihood that damage will arise is

[11] Note that this differs from Acemoglu and Lensman (2024) who assume that learning can be passive. A relaxation of this core assumption will be considered below.

updated by lowering the posterior probability that damage will arise. This is essentially an assumption that "no news is good news." More precisely, it is an assumption that a signal of harm reveals the state perfectly, while a signal there is no harm during a period involves the possibility of a false positive. Finally, observe that if there is no adoption of, say, B in period 1 (i.e., $I = 1$), then $\tilde{\mu}_B = \mu$ while $\tilde{\mu}_A = 0$; that is, if there is no indication in harm from A when it operates at "full" scale, then the probability that it is harmful in period 2 falls to 0. In contrast, if $I_1 = I_2 = \frac{1}{2}$ then $\tilde{\mu}_A = \tilde{\mu}_B = \frac{\mu}{2-\mu}$.

How might signals trigger policy intervention? Clearly, if one or both architectures advance and one or both receive a signal that it is harmful, this will be a trigger for intervention. When there is no adverse signal, either because neither architecture advances at $t = 1$ (and is therefore not adopted) or when both advance without an adverse signal despite adoption (leading to $\tilde{\mu}_j < \mu$ for both A and B), then there is no basis for intervention. It is possible, however, that intervention may be triggered by relatively "good news" if one architecture only advances but receives no adverse signal. In this case, the posterior probability of harm is lower for the architecture that has advanced, raising the question of whether intervention would be welfare-improving directed at the architecture that has not advanced.

Given this, suppose, without loss in generality, that A has advanced in period 1 (i.e., $q_A = 1$). This implies that there are three possible scenarios at the end of period 1 that would trigger potential interventions aimed at architecture A at the beginning of period 2:

1. (B harmful) $q_B = 1$ and $\delta_A = \delta_B = \delta$: B has advanced and both are known to be harmful;

2. (B harm uncertain) $q_B = 1$ and $\Pr[\delta_B = \delta] = \tilde{\mu}_B < 1$ and $\delta_A = \delta$: B has advanced but is not known to be harmful while A is known to be harmful;

3. (B has not advanced) $q_B = 0$ and $\delta_A = \delta$: B has not advanced and A is known to be harmful.

Note that it is also possible that $q_B = 1$ and $\delta_B = \delta$ and $\Pr[\delta_A = \delta] = \tilde{\mu}_A < 1$ where B has advanced and is known to be harmful while A's harm is uncertain. However, this is technically equivalent to Scenario 2 above (B harm uncertain), and so is omitted from consideration. In what follows, each policy option is considered for each of these four scenarios.

Before doing this, it is worthwhile noting another type of intervention for the scenario where A only advances and does not receive a signal of harm. Given that this scenario involves $I_1 = 1$, it is known with certainty that A is 'safe' (i.e., $\tilde{\mu}_A = 0$) whereas the probability that B is harmful remains at μ. In this situation, policymakers could preemptively enact regulations targeting B. For instance, regulators could prevent any potential harm from B

by either banning its adoption, banning further research on B or taxing B as if the harm would occur. In each case, $I_2 = 1$ (with all sectors adopting A) and $s_A = 1$ as a result. Expected social welfare would be $h(1)\int_0^1 2\Delta\, di + (1 - h(1))\int_0^1 \Delta\, di = (1 + h(1))\frac{1}{2}\Delta$ in each case. By contrast, if there was no intervention, if B advances, this leads to some value from its use but also a probability, μ, that there is harm resulting from that use. If μ and/or δ is relatively high, it may be worthwhile to intervene to 'not take a risk' on B. In what follows, this type of intervention is set aside, although it does highlight the continuing theme here that the trade-offs regarding intervention can be subtle.

5.2 Comparing Policy Options

Table 1 summarises the outcomes in terms of the allocation of scientists and the adoption of A and B architectures in period 2 under various scenarios. Note that endogenous outcomes are depicted with a 'hat' while policy requirements are hat-free. The endogenous items in blue are those that are socially optimal. In what follows, each outcome is described before turning to consider their rankings in terms of social welfare realised. In the appendix, expected social welfare for each option is derived and is the basis for the results to follow.

5.2.1 Ban on Adoption

This policy involves banning the adoption of any technology architecture for which it is known that $\delta_j = \delta$. If A is known to involve harm, its adoption is prohibited while, for B, prohibition only arises if it has advanced and adopted in period 1, as this is the only case where a signal of the harmfulness of adoption is generated and that signal realisation is that it is harmful.

Recalling that there can be no perfect signal that B is safe given that $1 - I_1 < 1$, B's harm remains uncertain in two cases and in those all scientific resources are allocated to B because there is no return from researching on A. Thus, with probability $h(1)$, B advances and, in so doing, competes with A at quality $Q_A(2) = \Delta$. Note, however, there remains a risk that B is harmful but this is not taken account in the adoption decision. In the one case where B is known to be harmful, the adoption of both A and B are barred so there is no research and total welfare is 0. The social welfare calculations are derived in the appendix.[12]

[12] Note that there is a potential time inconsistency issue. Because research on A is not prohibited, should it occur and should A advanced to $Q_A(2) = 2\Delta$, then it is possible that a policy-maker who has not committed to the ban may have an incentive to reverse the ban on A as $2\Delta > \delta$. If this reversal were anticipated, this would justify researching being conducted on the A architecture. This possibility is not evaluated here as it is assumed that the planner's policy implementations are time consistent.

Scenario	B harmful	B harm uncertain	B not advanced
$(q_A, q_B, \delta_A, \delta_B)$	$(1, 1, \delta, \delta)$	$(1, 1, \delta, \delta_B)$	$(1, 0, \delta, \delta_B)$
Adoption Ban	No Adoption of A or B	$I = \hat{s} = 0 \implies$ wp $h(1)$: B advances	$I = \hat{s} = 0 \implies$ wp $h(1)$: B advances
Research Prohibition	$s_A = s_B = 0$ $\hat{I} = \frac{1}{2}$	$s_A = 0$, $\hat{s}_B = 1 \implies$ wp $1 - h(1)$: $\hat{I} = \frac{1}{2}$ $h(1)$: $\hat{I} = \frac{1}{3}$	$s_A = 0$, $\hat{s}_B = 1 \implies$ wp $1 - h(1)$: $\hat{I} = 1$ $h(1)$: $\hat{I} = \frac{1}{2}$
Pigovian Tax	$\hat{s} = \frac{1}{2} \implies$ wp $h(\frac{1}{2})^2$: $\hat{I} = \frac{1}{2}$ $h(\frac{1}{2})(1 - h(\frac{1}{2}))$: $\hat{I} = 1$ $(1 - h(\frac{1}{2}))h(\frac{1}{2})$: $\hat{I} = 0$	$\hat{s} \geq 0 \implies$ wp $h(\hat{s})h(1 - \hat{s})$: $\hat{I} = \frac{2\Delta - \delta}{4\Delta - \delta}$ $h(\hat{s})(1 - h(1 - \hat{s}))$: $\hat{I} = \frac{2\Delta - \delta}{3\Delta - \delta}$ $(1 - h(\hat{s}))h(1 - \hat{s})$: $\hat{I} = 0$	$\hat{s} \geq 0 \implies$ wp $h(\hat{s})h(1 - \hat{s})$: $\hat{I} = \frac{2\Delta - \delta}{3\Delta - \delta}$ $h(\hat{s})(1 - h(1 - \hat{s}))$: $\hat{I} = 1$ $(1 - h(\hat{s}))h(1 - \hat{s})$: $\hat{I} = 0$
Ex Post Liability	$\hat{s} = \frac{1}{2} \implies$ wp $h(\frac{1}{2})^2$: $\hat{I} = \frac{1}{2}$ $h(\frac{1}{2})(1 - h(\frac{1}{2}))$: $\hat{I} = 1$ $(1 - h(\frac{1}{2}))h(\frac{1}{2})$: $\hat{I} = 0$	$\hat{s} \geq 0 \implies$ wp $h(\hat{s})h(1 - \hat{s})$: $\hat{I} = \frac{2\Delta - \delta}{4\Delta - (1 + \tilde{\mu})\delta}$ $h(\hat{s})(1 - h(1 - \hat{s}))$: $\hat{I} = \frac{2\Delta - \delta}{3\Delta - \delta + \tilde{\mu}\Delta}$ $(1 - h(\hat{s}))h(1 - \hat{s})$: $\hat{I} = 0$	$\hat{s} \geq 0 \implies$ wp $h(\hat{s})h(1 - \hat{s})$: $\hat{I} = \frac{2\Delta - \delta}{3\Delta - \delta + \mu\Delta}$ $h(\hat{s})(1 - h(1 - \hat{s}))$: $\hat{I} = 1$ $(1 - h(\hat{s}))h(1 - \hat{s})$: $\hat{I} = 0$

Table 1: **Research and Adoption Outcomes**

5.2.2 Prohibition of Research

This policy involves prohibiting research from advancing any technology architecture for which it is known that $\delta_j = \delta$. If A is known to involve harm, research on A is prohibited while, for B, prohibition could only arise if it has advanced and adopted in period 1, where a prohibition is put in place if the resulting signal is that it is harmful.

In this case, where B is still not known to be harmful, all scientific resources are allocated to B. If B research is successful, then this creates competition for A and adoption moves towards a greater number of sectors using B. That competition is beneficial as any sector adopting A is welfare-reducing as $\Delta < \delta$. When B is known to be harmful, there is no further research. In this case, adopting either technology is welfare reducing but because there is no ban on adoption that occurs in all sectors.

We are now in a position to compare the policies of a ban on adoption and a prohibition on research. We can establish the following.

Proposition 3 *A prohibition of research always results in lower expected social welfare than* *a ban on adoption.*

The intuition is straightforward and can be seen in Table 1. An adoption ban leads to the same research outcome as a research prohibition but ensures that there is no adoption of A as $Q_A(2) = \Delta$ which would be strictly welfare reducing. Such adoption can still occur when there is a research prohibition only.

5.2.3 Pigouvian Taxation

Pigouvian taxation involves levying a sector-specific charge, $\tau_{i,j} = -E_{i,j} = \eta_{i,j}\delta$, on each sector adopting a technology j that is *known* to be harmful. As it is assumed that A is known to be harmful, then a full internalisation of the externality would imply a tax of $\tau_{i,A} = -E_{i,A} = \eta_{i,A}\delta$ for all $i < \hat{I}$. Note that while this tax would eliminate the adoption of A if $Q_A(2) = \Delta$, A will still be adopted if $Q_A(2) = 2\Delta$. This plays an important role in driving the social and private incentives to research in period 2 along the A path.

Once again, the full social welfare outcomes are stated in the appendix. It is instructive to compare the outcomes under Pigouvian taxation with those from a ban on adoption. It is often the case that Pigouvian taxation that fully internalises the external harm for a decision-maker, here a sector adopting a harmful technology, leads to higher social welfare outcomes than a ban on the decision that may lead to harm. This certainly is the case when B is also known to harmful. In that situation, the Pigouvian tax leaves open the possibility of adoption should there be further advances in one or both architectures and so expected social welfare is higher than zero; the level that results from a complete ban on the adoption of both architectures.

This simple intuition breaks down, however, when there continues to be uncertainty regarding whether the B technology is harmful or not. In this case, while the A technology is only adopted if that technology advances in period 2 (as $\Delta < \delta < 2\Delta$), the B technology might be adopted even if $\Delta < \tilde{\mu}\delta$. In this case, it is possible that in certain states expected social welfare may be negative even under a Pigouvian tax. Nonetheless, it remains the case that expected social welfare at the beginning of period 2 is higher under a Pigouvian tax than a ban on adoption.

Proposition 4 *A Pigouvian tax always results in a (weakly) higher expected social welfare than a ban on adoption.*

The reason that a Pigouvian tax socially dominates a ban on technology adoption even when B harm is uncertain is because the B technology might be adopted even when A has been banned. When $\hat{s} \to 0$ under a Pigouvian tax, expected social welfare is the same as under a ban on adoption. However, it is possible that in some circumstances, $\hat{s} > 0$ under a Pigouvian tax, leading to some A adoption that is superior to B adoption for some sectors if $2\Delta - \delta > \Delta - \tilde{\mu}\delta$.

The potential social inefficiency from a Pigouvian tax arises because B adopters do not internalise the *expected* harm from their adoption. This insight leads to the following result that demonstrates that a Pigouvian tax that fully internalises the harm from A adopters leads to too little A adoption and too few scientists allocated to advancing A further.

Proposition 5 *Under a Pigouvian tax of $\tau_{i,A} = \eta_{i,A}\delta$ for all i adopting A, when the harmfulness of B remains unknown, increasing the allocation of scientists to the A architecture would raise expected social welfare.*

The proof is a straightforward examination of the expected social welfare calculations in the Appendix. Intuitively, when the harm to B remains unknown, its expected harm of $\tilde{\mu}\delta$ or $\mu\delta$ as the case may be, is not taken into account in either the adoption of B by sectors or the returns to B research. If it were taken into account, both would be reduced. Thus, from a social perspective, a Pigouvian tax whereby A adopters internalised fully the known harm would result in too little research being directed towards improving A. One mechanism that could achieve this would be to lower $\tau_{i,A}$.

A natural question following this result is what the optimal Pigouvian tax would be. A precise characterisation of this would be complex and really only establish that a second-best outcome could be achieved with a tax that does not fully internalise A's external harm; that is, is lower than $\eta_{i,A}\delta$. The real issue is that the optimal tax needs to be contingent upon the realised level of both architectures, which means it must be determined ex-post after the realisation of research outcomes during period 2. The following result characterises the optimal (ex post) tax on A:

Proposition 6 *Suppose that B is not known to be harmful at the beginning of period 2 and that $h(s_j) = s_j^a$ ($a \in (0,1)$). The optimal Pigouvian taxes following the realisation of period 2 research outcomes are as follows:*

1. *if $Q_A(2) = Q_B(2) = 2\Delta$, $\tau^*_{i,A} = \eta_{i,j} \max\{\frac{\Delta(1-3\tilde{\mu})\delta}{\Delta-\tilde{\mu}}\delta, 0\}$;*
2. *if $Q_A(2) = 2\Delta$ and $Q_B(2) = \Delta$, $\tau^*_{i,A} = \eta_{i,j} \max\{\frac{\Delta(1-2\tilde{\mu})\delta}{\Delta-\tilde{\mu}}\delta, 0\}$; and*
3. *if $Q_A(2) = \Delta$ and/or $Q_B(2) = 0$, $\tau^*_{i,A} = \eta_{i,A}\delta$.*

The proof follows directly from maximising expected social welfare as calculated in the appendix and noting that the assumption on $h(.)$ guarantees that the scientist allocation is optimal as per Lemma 1. In its absence, the difference between the socially optimal and regulated outcome will reflect the earlier results from Lemma 1.

These taxes are optimal because they are such that the adoption of A is optimal ex post (that is, so that $\hat{I} = \frac{Q_A(2)-\delta}{Q_A(2)-\delta+Q_B(2)-\tilde{\mu}\delta}$). Note that they may be contingent on the posterior probability, $\tilde{\mu}$, which itself depends on whether any B adoption occurred in period 1. The higher is $Q_B(2)$, the lower is $\tau^*_{i,A}$. This is because it is when B is more competitive that, at the margin, A's competitiveness needs to be strengthened the most.

Interestingly, the proposition demonstrates that there are situations when it may be optimal not to have a tax at all. That is, if either $2\Delta - \delta \geq \Delta - \tilde{\mu}\delta \Leftrightarrow \tilde{\mu} \geq \frac{\delta-\Delta}{\delta}$ or $\tilde{\mu} \geq \frac{1}{3}$ depending on the case. In these situations, the optimal policy would be a subsidy to A rather than a tax. This reflects the observation made earlier that if μ is sufficiently high, then the socially optimal insurance against harm from a technology is to develop a technology sufficiently advanced to 'pay' for that harm rather than prevent the adoption of harmful technology per se.

5.3 Ex post liability

When the Pigouvian tax can be applied ex post and tailored to the realisation of research outcomes, the socially optimal outcome can be produced (at least when $h(s_j) = s_j^a$). There may be practical difficulties in doing this if the precise research outcomes cannot be observed easily by the social planner. However, the final policy option of ex post liability is designed to allow for some degree of tailoring to realised outcomes and, therefore, may result in higher expected social welfare than Pigouvian taxation.

Ex post liability involves imposing a penalty on technology adopters if there is adoption that results in realised harm. While in the model presented here, this possibility does not change research and adoption decisions in period 1 (that is, a penalty may be imposed, but its expectation does not alter the relative allocation of scientific resources nor the adoption decisions), it does impact on period 2 decisions both in terms of realised outcomes prior to that point (akin to the impact of a Pigouvian tax) but also in expectation of potential harm (unlike a Pigouvian tax). Therefore, as is summarised in Table 1, in numerous scenarios, the expectation of a penalty ex post generates a socially optimal adoption decision.

What restricts a socially optimal adoption decision is that it is assumed here that adopters have limited liability in that their realised penalty cannot exceed the realised surplus when a technology is adopted. Note that surplus is the liability metric because profits accrue to both adopters of a technology and also providers of the technology (i.e., scientists). It is assumed that both are liable for any realised harm.[13] For instance, if technology adoption results in surplus of $\eta_{i,j}2\Delta$, this can always fund a penalty of $\eta_{i,j}\delta$. Hence, when both technologies generate this surplus, the limited liability constraint is not binding, and adoption is socially optimal.

However, if the adoption of a technology results in a surplus of $\eta_{i,j}\Delta$, this also defines the maximum penalty, which is less than $\eta_{i,j}\delta$. The limited liability constraint binds, and thus, there may be too much adoption prior to the resolution of uncertainty regarding harm. The only time, however, when this leads to sub-optimal adoption is when adopting technology A generates a surplus of $\eta_{i,A}2\Delta$ while B generates a surplus of $\eta_{i,B}\Delta$. In this case, the limited liability constraint applies for B but not for A, resulting in too little adoption of A feeding into too few scientific resources devoted to advancing A. If liability were unlimited, this distortion would not arise, and adoption would be socially optimal.

Nonetheless, even limited liability pushes adoption closer to a socially optimal level than does Pigouvian taxation. For that reason, the following result can be demonstrated:

Proposition 7 *An ex post liability regime always results in (weakly) higher expected social welfare than a Pigouvian tax.*

Thus, this completes the comparison of each of the four policy options. Put simply, the more a policy instrument can adjust in its application to the realisation of uncertainty along both the harm and research outcome dimensions, the closer the outcome will be to what would be socially optimal. This favours policy options that are ex-post in nature and only determined in their application following the realisation of uncertainty but where, in anticipation of that adjustment, it impacts the expectations of the relevant decision-makers – scientists and technology adopters.[14]

6 Extensions

Various assumptions were made in deriving the above results. In this section, these assumptions are re-examined, and the implications of generalising them are considered.

[13] This does not necessarily reflect how tort law might be applied, which may only find one of these agent types liable, in which case the liability constraint will bind more strongly.

[14] Guerreiro et al. (2023) examines regulatory options with respect to AI and finds that a liability regime is socially optimal in their context if liability is unlimited. They examine different margins of AI adoption than the focus on the direction of technological change here.

6.1 Learning from Research

The model thus far assumes that signals regarding harm arise from AI adoption. That is, harm levels are surfaced through direct marketplace testing or 'learning by doing.' Harms can also be signalled through research or what Gans (2024) calls 'lab learning.' This is the type of learning considered by Acemoglu and Lensman (2024). The analogue to the earlier updating formula would be that if there is no signal of harm following a research period, the posterior probability of harm becomes:

$$\tilde{\mu}_A = \frac{\mu(1-s)}{\mu(1-s)+(1-\mu)}$$

$$\tilde{\mu}_B = \frac{\mu s}{\mu s+(1-\mu)}$$

Thus, if all research is devoted to one path, the signal of harm will be perfect; otherwise, there is some learning, but uncertainty remains.

From a policy perspective, the post-research probabilities of harm can be taken into account when deciding whether to adopt technologies should they have been developed. As harm only arises from adoption, this generates an incentive to conduct research in order to evaluate harm and potentially avoid incurring any harm. This learning and the option it affords are valuable for regulatory interventions that are contingent upon those signals, such as banning adoption (or further research) and a Pigouvian tax. For ex post liability, however, the new information can be used in decentralised scientist and adoption decisions and will reduce errors.

Apart from these details, however, having lab learning rather than learning by doing changes the overall regulatory picture along the lines outlined by Gans (2024) in that it potentially introduces a precautionary motive to any AI adoption and would prioritise research to surface harms should that be possible.

6.2 Higher and Lower Damage

If harm occurs, it has been assumed that $\delta \in [\Delta, 2\Delta)$. Within this range, it is socially optimal to adopt technology along a path if it is sufficiently advanced (i.e., $q_j = 2$) and not otherwise. It is this assumption that generated a potential incentive to continue research along a path even if the risk of harm had increased over time.

There are two changes that may arise if this assumption is relaxed. If $\delta < \Delta$, then it is never optimal not to adopt a technology, even if it is known to be harmful. Thus, there is no case for intervention to prohibit adoption, although there remains an incentive to push adoption towards a less harmful path should it exist. Otherwise, the incentives to choose regulatory instruments that internalise externalities and risk remain.

If $\delta > 2\Delta$, then it is never optimal to adopt a technology known to be harmful. This strengthens the case for banning adoption and/or research along a harmful path, as it is no longer the case that advancing the technology sufficiently can outweigh the costs it imposes. Thus, this paper could be interpreted as favouring policies that allow the pricing-in of externalities into private decisions so long as the harm evaluated is not too high. If that harm is known to be substantial, this bolsters the case for prohibitions as regulatory instruments.

6.3 Damage that Scales

The final assumption worth examining is that the extent of damage is independent of the quality of the technology adoption. By contrast, suppose that $E_{i,j}(t) = \eta_{i,j}\delta_j Q_j(t)$; that is, damage scales with the quality of the technology. This is the main case considered by Acemoglu and Lensman (2024).[15] This implies that total social welfare at a given time is:

$$\int_0^I \eta_{i,A}(1-\delta_A)Q_A(2)di + \int_I^1 \eta_{i,B}(1-\delta_B)Q_B(2)di$$

The optimum sectoral adoption threshold satisfies:

$$I^*(t) = \frac{Q_A(t)(1-\delta_A)}{Q_A(t)(1-\delta_A) + Q_B(t)(1-\delta_B)}$$

Importantly, this implies that it is optimal to adopt a technology with $Q_j(t) > 0$ regardless of its quality if and only if $\delta_A < 1$.

It can readily be seen that this specification simplifies the analysis of the model akin to the cases of higher ($\delta_j > 1$) and lower ($\delta_j < 1$) damage considered in the previous subsection. However, it does not allow the more complex trade-offs that arise in the intermediate case, which is the focus of this paper.

However, this specification could open up various policy commitment issues that may be pursued in future work. For instance, for δ always less than 1, a policy-maker may want to commit to not adopting a technology with known harm so as to encourage scientific research on the other technology path. However, when policymakers learn about the level of harm, they may be unable to commit to de-adopting the technology. Thus, there may be a time inconsistency issue that, in turn, makes certain decisions irreversible. As Gans (2024) argues, irreversibility can change the value of learning about harm. Examining this would require a

[15] The case without such scaling is considered in the main model here is in an appendix in their work.

more detailed model of the regulator than is provided here, and so it is left for future work.

7 Conclusion

This paper has examined the complex trade-offs involved in regulating the direction of technological innovation, with a particular focus on situations where potential harms from new technologies are uncertain. The analysis reveals several key insights. First, the socially optimal allocation of scientific resources between competing technological paths is not always straightforward. While diversification can provide insurance against potential harm, there are also cases where concentrating resources on advancing a leading technology may be preferable, even if that technology carries some risk of harm. Second, market forces alone tend to result in an inefficiently high concentration of research effort on leading technological paths. This stems from scientists failing to account for the negative externality their choice imposes on the probability of success for the lagging path. Third, when regulating in the face of uncertainty about potential harms, ex post policy instruments that can adjust to realized outcomes tend to outperform ex ante prohibitions or restrictions. Specifically, the analysis suggests that ex post liability regimes are likely to produce better outcomes than Pigouvian taxes, which in turn outperform bans on adoption or research. Fourth, the optimal regulatory approach depends critically on the magnitude of potential harm relative to the benefits of technological progress. For very high levels of harm, prohibitions may become optimal, while for lower levels, instruments that allow for pricing-in of risk are preferable. Finally, there is an important distinction between learning about harms through research versus through adoption. The possibility of ”lab learning” introduces additional complexity to the optimal regulatory strategy.

These findings have important implications for current debates surrounding the regulation of emerging technologies like AI. They suggest that policymakers should be cautious about implementing heavy-handed ex ante restrictions on research or adoption paths. Instead, the focus should be on developing robust mechanisms for ongoing assessment of potential harms and flexible policy instruments that can adjust as uncertainty is resolved.

However, several important questions remain for future research. These include exploring how different liability regimes might be structured to optimize incentives, examining the implications of strategic behaviour by firms or researchers in anticipation of future regulation, and investigating how international coordination (or lack thereof) impacts the efficacy of different regulatory approaches.

8 Appendix

8.1 Proof of Lemma 1

Suppose $s^* = \frac{1}{2}$, then $h'(s^*) = h'(1-s^*)$ which implies that $\frac{h(\hat{s})/\hat{s}}{h(1-\hat{s})/(1-\hat{s})}$ or $\hat{s} = \frac{1}{2}$. If $q_A = q_B = q$, then $\frac{h'(s^*(q,q))}{h'(1-s^*(q,q))} = \frac{h(\hat{s}(q,q))/\hat{s}(q,q}{h(1-\hat{s}(q,q))/(1-\hat{s}(q,q))} = 1$. In this case, $s^* = \frac{1}{2}$.

Next, note that $\lim_{s\to 0}\left(\frac{h'(s)}{h'(1-s)} - \frac{h(s)/s}{h(1-s)/(1-s)}\right) > 0$ and $\lim_{s\to 1}\left(\frac{h'(s)}{h'(1-s)} - \frac{h(s)/s}{h(1-s)/(1-s)}\right) < 0$ by the assumed Inada conditions on $h(.)$. If $\frac{h'(s)}{h(s)/s}$ is decreasing in s, $\frac{h'(s)}{h'(1-s)}$ and $\frac{h(s)/s}{h(1-s)/(1-s)}$ cross at exactly one point, which, as already demonstrated, is where $s = \frac{1}{2}$.

Note that $q_A \neq q_B$ if $\{q_A, q_B\} = \{1, 0\}$ or $\{0, 1\}$. Consider the case where $\{q_A, q_B\} = \{1, 0\}$. Note that $\frac{1-\frac{1}{3}h(s)}{2-\frac{1}{3}h(1-s)}$ has the following properties: (i) it is decreasing in s; (ii) as $s \to 0$, this becomes $\frac{1}{2-\frac{1}{3}h(1)} < 1$; (iii) as $s \to 1$, this becomes $\frac{1-\frac{1}{3}h(1)}{2} < \frac{1}{2-\frac{1}{3}h(1)}$; and (iv) that $\frac{\partial \frac{h'(s)}{h'(1-s)}}{\partial s} < \frac{\partial \frac{1-\frac{1}{3}h(s)}{2-\frac{1}{3}h(1-s)}}{\partial s}$, as $\frac{h'(1)}{h'(0)} = 0 < \frac{1-\frac{1}{3}h(1)}{2}$. This implies that $\frac{1-\frac{1}{3}h(s^*(1,0))}{2-\frac{1}{3}h(1-s^*(1,0))} < 1$ and so $s^* > \frac{1}{2}$ and $\frac{h(s^*)/s^*}{h(1-s^*)/(1-s^*)} > \frac{1-\frac{1}{3}h(s^*(1,0))}{2-\frac{1}{3}h(1-s^*(1,0))}$. Condition (iv) above then implies that $\hat{s}(1,0) > s^*(1,0)$. An analogous argument holds where $s^* < \frac{1}{2}$.

8.2 Proof of Proposition 1

First, some preliminary calculations. Note that: $v_S(0,0,\delta,\delta) = 0$, $v_S(1,0,\delta,\delta) = 0$, $v_S(1,1,\delta,\delta) = 0$, $v_S(2,0,\delta,\delta) = \frac{1}{2}(2\Delta - \delta)$, $v_S(2,1,\delta,\delta) = \frac{1}{2}(2\Delta - \delta)$ and $v_S(2,2,\delta,\delta) = \frac{3}{4}(2\Delta - \delta)$ while $v_S(0,0,0,0) = 0$, $v_S(1,0,0,0) = \frac{1}{2}\Delta$, $v_S(1,1,0,0) = \frac{3}{4}\Delta$, $v_S(2,0,0,0) = \Delta$, $v_S(2,1,0,0) = \frac{7}{6}\Delta$ and $v_S(2,2,0,0) = \frac{3}{2}\Delta$. Moreover, when the realisation of externalities is different we have: $v_S(0,0,\delta,0) = 0$, $v_S(1,0,\delta,0) = 0$, $v_S(1,1,\delta,0) = \frac{1}{2}\Delta$, $v_S(2,0,\delta,0) = \frac{1}{2}(2\Delta - \delta)$, $v_S(2,1,\delta,0) = \frac{1}{2}(\Delta + \frac{(2\Delta-\delta)^2}{3\Delta-\delta})$, $v_S(2,2,\delta,0) = \frac{1}{2}\left(2\Delta - \frac{(2\Delta-\delta)^2}{4\Delta-\delta}\right)$, $v_S(1,0,0,\delta) = \frac{1}{2}\Delta$, $v_S(2,0,0,\delta) = \Delta$ and $v_S(2,1,0,\delta) = \Delta$.

Using these, the following can be calculated:

1. $q_A = q_B = 1$: $\mathbb{E}[v_s(2,2)] = \frac{24\Delta^2+\delta^2(4-\mu)\mu-6\delta\Delta(2\mu+1)}{4(4\Delta-\delta)}$, $\mathbb{E}[v_s(1,1)] = \frac{1}{4}\Delta(1-\mu)(\mu+3)$ and $\mathbb{E}[v_s(2,1)] = \mathbb{E}[v_s(1,2)] = \frac{3\delta^2\mu-\delta\Delta(\mu(\mu+7)+7)+3\Delta^2(7-\mu)}{6(3\Delta-\delta)}$;

2. $q_A = q_B = 0$: $\mathbb{E}[v_s(1,1)] = \frac{1}{4}\Delta(1-\mu)(\mu+3)$, $\mathbb{E}[v_s(1,0)] = (1-\mu)\frac{1}{2}\Delta$ and $\mathbb{E}[v_s(0,0)] = 0$; and

3. $q_A = 1, q_B = 0$: $\mathbb{E}[v_s(2,1)] = \frac{3\delta^2\mu-\delta\Delta(\mu(\mu+7)+7)+3\Delta^2(7-\mu)}{6(3\Delta-\delta)}$, $\mathbb{E}[v_s(2,0)] = \Delta - \frac{\delta\mu}{2}$, $\mathbb{E}[v_s(1,1)] = \frac{1}{4}\Delta(1-\mu)(\mu+3)$, and $\mathbb{E}[v_s(1,0)] = (1-\mu)\frac{1}{2}\Delta$.

Given this, we have:

$$\frac{h'(s)}{h'(1-s)} = \frac{\Delta(1-\mu)\left(h(s)(\delta+5\delta\mu-3(\Delta+3\Delta\mu))-3(\mu+1)(3\Delta-\delta)\right)}{\Delta(1-\mu)h(1-s)(\delta+5\delta\mu-3(\Delta+3\Delta\mu))+6(3\Delta-\delta)(\Delta(\mu+1)-\delta\mu)}$$

Note that at $s=\frac{1}{2}$, the numerator of the left-hand side is less than the denominator if:

$$-3\Delta(1-\mu)(\mu+1)(3\Delta-\delta) < 6(3\Delta-\delta)(\delta\mu-\Delta(\mu+1)) \implies -(1-\mu)(1+\mu)\Delta < 2(\delta\mu-\Delta(1+\mu))$$

Note that at $\delta=\Delta$, this simplifies to $-(1-\mu)(1+\mu) < -2$ which holds for $\mu<1$. Thus, by the concavity of $h(.)$, $s^* > \frac{1}{2}$. It is also straightforward to determine that the left-hand side of the first order condition is increasing in δ, implying that s^* is decreasing in δ. The opposite is true for Δ.

With regard to μ, the left-hand side of the first-order condition is increasing in μ for μ sufficiently low and decreasing thereafter. Specifically, the derivative of the left-hand side as $\mu \to 1$ is equal to $\frac{\Delta((2\Delta-\delta)h(s)-(3\Delta-\delta))}{(3\Delta-\delta)(2\Delta-\delta)}$ which is clearly negative.

8.3 Proof of Proposition 3

The only difference between the private and social returns happens when both advance and B does not have an externality. In this case, a private scientist is comparing:

$$\frac{h(s)}{s}\left((1-\mu)h(1-s)\frac{(2\Delta-\delta)^2}{2(3\Delta-\delta)}+(1-(1-\mu)h(1-s))\frac{2\Delta-\delta}{2}\right)$$

and

$$\frac{h(1-s)}{1-s}(1-\mu)\left((1-h(s))\frac{\Delta}{2}+h(s)\frac{\Delta^2}{2(3\Delta-\delta)}\right)$$

whereas the social return is:

$$\begin{aligned} h'(s)\left(\int_0^1(1-i)(2\Delta-\delta)di-h(1-s)(1-\mu)\left(\int_{\frac{2\Delta-\delta}{3\Delta-\delta}}^1(1-i)(2\Delta-\delta)di+\int_0^{\frac{2\Delta-\delta}{3\Delta-\delta}}i\Delta di\right)\right)\\ =h'(1-s)\left((1-\mu)\int_0^1 i\Delta di-h(s)(1-\mu)\left(\int_{\frac{2\Delta-\delta}{3\Delta-\delta}}^1(1-i)(2\Delta-\delta)di+\int_0^{\frac{2\Delta-\delta}{3\Delta-\delta}}i\Delta di\right)\right)\end{aligned}$$

Compared with the private return allocation equation:

$$\begin{aligned}
&\frac{h(s)}{s}\left(\int_0^1 (1-i)(2\Delta-\delta)di - h(1-s)(1-\mu)\left(\int_{\frac{2\Delta-\delta}{3\Delta-\delta}}^1 (1-i)(2\Delta-\delta)di + \int_0^{\frac{2\Delta-\delta}{3\Delta-\delta}} i\Delta di\right)\right) \\
&\quad= \frac{h(1-s)}{1-s}\left((1-\mu)\int_0^1 i\Delta di - h(s)(1-\mu)\left(\int_{\frac{2\Delta-\delta}{3\Delta-\delta}}^1 (1-i)(2\Delta-\delta)di + \int_0^{\frac{2\Delta-\delta}{3\Delta-\delta}} i\Delta di\right)\right)
\end{aligned}$$

It can be seen that these only differ with respect to $\frac{h'(s)}{h'(1-s)}$ and $\frac{(1-s)h(s)}{sh(1-s)}$.

8.4 Social Welfare under Each Policy Option

Here, for completeness, social welfare under each policy option is calculated reflecting the main comparator outcomes in Table 1. The results from Propositions 4 - 7 follow from these calculations as outlined in the text.

8.4.1 Ban on Adoption

1. (B harmful) $q_B = 1$ and $\delta_A = \delta_B = \delta$: No research would be undertaken in period 2 (or more specifically, any research would be inconsequential), and social welfare would be $V^{NoAd}(1, 1, \delta, \delta) = 0$.

2. (B harm uncertain) $q_B = 1$ and $\delta_A = \delta$ and δ_B remains uncertain with posterior probability of $\tilde{\mu}_B = \frac{\mu}{2-\mu}$ as $\hat{I}_1 = \frac{1}{2}$. As the adoption of A is prohibited, then there will be no research devoted to extending it at $t = 2$ as no scientist can earn a return on any advance. Therefore, s_B will be as high as possible, resulting in $\hat{s} = 0$. Expected social welfare will, therefore, be as $\hat{I}_2 = 0$:

$$\begin{aligned}
\mathbb{E}[V^{NoAd}(1, 1, \delta, \delta_B)] &= h(1)\int_0^1 i(2\Delta - \tilde{\mu}_B\delta)\, di + (1 - h(1))\int_0^1 i(\Delta - \tilde{\mu}_B\delta)\, di \\
&= \tfrac{1}{2}\left((1 + h(1))\Delta - \delta\tilde{\mu}_B\right)
\end{aligned}$$

3. (B has not advanced) $q_B = 0$ and $\delta_A = \delta$: Again this implies that $s_A = 0$ and so $s_B = 1$. Thus, $\mathbb{E}[V^{NoAd}(1, 0, \delta, \delta_B] = h(1)\left(\mu\frac{1}{2}(\Delta - \delta) + (1 - \mu)\frac{1}{2}\Delta\right) = h(1)\frac{1}{2}(\Delta - \mu\delta)$.

8.4.2 Prohibition on Research

1. (B harmful) $q_B = 1$ and $\delta_A = \delta_B = \delta$: No research would be undertaken in period 2; however, because both architectures have advanced, $\hat{I}_2 = \frac{1}{2}$ and so social welfare will be $V^{NoRes}(1, 1, \delta, \delta) = \frac{1}{2}(\Delta - \delta) < 0$.

2. (B harm uncertain) $q_B = 1$ and $\delta_A = \delta$ and δ_B remains uncertain with a posterior probability of $\tilde{\mu}_B$: The prohibition on A research implies that all scientists will be allocated to research on B. Thus, $s_B = 1$ (or equivalently), $\hat{s} = 0$. Expected social welfare will, therefore, be:

$$\begin{aligned}\mathbb{E}[V^{NoRes}(1,1,\delta,\delta_B)] &= (1-h(1))\left(\int_0^{\frac{1}{2}}(1-i)(\Delta-\delta)\,di + \int_{\frac{1}{2}}^1 i(\Delta-\tilde{\mu}_B\delta)\,di\right)\\ &+ h(1)\left(\int_0^{\frac{1}{3}}(1-i)(\Delta-\delta)\,di + \int_{\frac{1}{3}}^1 i(2\Delta\tilde{\mu}_B-\delta)\,di\right)\\ &= \frac{1}{72}(54\Delta - 27(1+\tilde{\mu}_B)\delta + h(1)(\delta(7-5\tilde{\mu}_B)+30\Delta))\end{aligned}$$

3. (B has not advanced) $q_B = 0$ and $\delta_A = \delta$: If research on A is prohibited, then all scientists will research on B. Note, however, as the adoption of A is not prohibited, and externalities are not internalised, then $\hat{I}(2) = \frac{1}{2}$ if there is an advance in B; otherwise, all sectors continue to adopt A. It is clear that this involves lower social welfare than prohibiting the adoption of A as any use of A lowers social welfare given that $\Delta < \delta$. Social welfare is:

$$\begin{aligned}\mathbb{E}[V^{NoRes}(1,0,\delta,\delta_B)] &= (1-h(1))\int_0^1(1-i)(\Delta-\delta)\,di\\ &+ h(1)\left(\int_0^{\frac{1}{2}}(1-i)(\Delta-\delta)\,di + \int_{\frac{1}{2}}^1 i(\Delta-\mu\delta)\,di\right)\\ &= \tfrac{1}{2}(\Delta-\delta) + h(1)\frac{2\Delta+(1-3\mu)\delta}{8}\end{aligned}$$

8.4.3 Pigouvian Tax

1. (B harmful) $q_B = 1$ and $\delta_B = \delta$: In this case, a tax of $\tau_B = E_{i,B}(2) = -\eta_{i,B}\delta$ is imposed on the adoption of B leading to an ex post optimal adoption of technologies with $\hat{I}(2) = \frac{1}{2}$ if both A and B advance to $Q_j(2) = 2\Delta$, $\hat{I}(2) = 1$ if only A advances, $\hat{I}(2) = 0$ if only B advances and no adoption if neither advance. Given this, the decentralised allocation of scientists will be $\hat{s} = \frac{1}{2}$, which is also the socially optimal

allocation. In this case, social welfare is:

$$\begin{aligned}\mathbb{E}[V^{Tax}(1,1,\delta,\delta)] =& h(\tfrac{1}{2})^2\left(\int_0^{\frac{1}{2}}(1-i)(2\Delta-\delta)\,di+\int_{\frac{1}{2}}^1 i(2\Delta-\delta)\,di\right)\\ &+h(\tfrac{1}{2})(1-h(\tfrac{1}{2}))\left(\int_0^1(1-i)(2\Delta-\delta)di+\int_0^1 i(2\Delta-\delta)di\right)\\ =& h(\tfrac{1}{2})(4-h(\tfrac{1}{2}))\tfrac{1}{4}(2\Delta-\delta)\end{aligned}$$

2. (B harm uncertain) $q_B=1$ and δ_B remains uncertain with posterior probability of $\tilde{\mu}$: Social welfare is:

$$\begin{aligned}\mathbb{E}[V^{Tax}(1,1,\delta,\delta)] =& h(1-s)h(s)\left(\int_{\frac{2\Delta-\delta}{4\Delta-\delta}}^1 i(2\Delta-\delta\tilde{\mu}_B)\,di+\int_0^{\frac{2\Delta-\delta}{4\Delta-\delta}}(1-i)(2\Delta-\delta)\,di\right)\\ &+h(1-s)(1-h(s))\int_0^1 i(2\Delta-\tilde{\mu}\delta)\,di\\ &+(1-h(1-s))h(s)\left(\int_0^{\frac{2\Delta-\delta}{3\Delta-\delta}}(1-i)(2\Delta-\delta)\,di+\int_{\frac{2\Delta-\delta}{3\Delta-\delta}}^1 i(\Delta-\tilde{\mu}\delta)\,di\right)\\ &+(1-h(s))(1-h(1-s))\int_0^1 i(\Delta-\tilde{\mu}\delta)\,di\end{aligned}$$

3. (B has not advanced) $q_B=0$ while $\tilde{\mu}=\mu$: Research potentially occurs on both paths in period 2. If neither advances, the social welfare (and private return) will be 0 as A will not be adopted under Pigouvian taxation. If A advances while B does not, then $\hat{I}(2)=1$ and social welfare (and A return) is $\frac{1}{2}(2\Delta-\delta)$. If B advances while A does not, then $\hat{I}(2)=0$ and the B return is $\frac{1}{2}\Delta$. Finally, if both advance, then $\hat{I}(2)=\frac{2\Delta-\delta}{3\Delta-\delta}$ with $v_A(2,1)=\int_0^{\frac{2\Delta-\delta}{3\Delta-\delta}}\left((1-i)(2\Delta-\delta)-i\Delta\right)\,di=\frac{(2\Delta-\delta)^2}{2(3\Delta-\delta)}$ and $v_B(2,1)=\int_{\frac{2\Delta-\delta}{3\Delta-\delta}}^1\left(i\Delta-(1-i)(2\Delta-\delta)\right)\,di=\frac{\Delta^2}{2(3\Delta-\delta)}$. At the beginning of period 2, $\hat{s}(2)$ will equate the average returns to scientists researching on each path; that is:

$$\frac{h(\hat{s}(2))}{\hat{s}(2))}\left(h(1-\hat{s}(2))\int_0^{\frac{2\Delta-\delta}{3\Delta-\delta}}\left((1-i)(2\Delta-\delta)-i\Delta\right)\,di+(1-h(1-\hat{s}(2)))\int_0^1(1-i)(2\Delta-\delta)di\right)$$

$$=\frac{h(1-\hat{s}(2))}{1-\hat{s}(2))}\left(h(\hat{s}(2))\int_{\frac{2\Delta-\delta}{3\Delta-\delta}}^1\left(i\Delta-(1-i)(2\Delta-\delta)\right)\,di+(1-h(\hat{s}(2)))\int_0^1 i\Delta di\right)$$

or

$$\frac{h(\hat{s}(2))/\hat{s}(2)}{h(1-\hat{s}(2))/(1-\hat{s}(2))} = \frac{\int_0^1 i\Delta\, di - h(\hat{s}(2))\left(\int_{\frac{2\Delta-\delta}{3\Delta-\delta}}^1 (1-i)(2\Delta-\delta)\, di + \int_0^{\frac{2\Delta-\delta}{3\Delta-\delta}} i\Delta\, di\right)}{\int_0^1 (1-i)(2\Delta-\delta) di - h(1-\hat{s}(2))\left(\int_{\frac{2\Delta-\delta}{3\Delta-\delta}}^1 (1-i)(2\Delta-\delta)\, di + \int_0^{\frac{2\Delta-\delta}{3\Delta-\delta}} i\Delta\, di\right)}$$

Thus, at the beginning of period 2, expected social welfare is:

$$\begin{aligned}
\mathbb{E}[V^{Tax}(1,0,\delta,\delta_B)] =& h(\hat{s}(2))h(1-\hat{s}(2))\left(\int_0^{\frac{2\Delta-\delta}{3\Delta-\delta}} (1-i)(2\Delta-\delta)\, di + \int_{\frac{2\Delta-\delta}{3\Delta-\delta}}^1 i(\Delta-\mu\delta)\, di\right) \\
&+ h(\hat{s}(2))(1-h(1-\hat{s}(2)))\int_0^1 (1-i)(2\Delta-\delta) di \\
&+ (1-h(\hat{s}(2)))h(1-\hat{s}(2))\int_0^1 i(\Delta-\mu\delta) di + (1-h(1-s))(1-h(s))0 \\
=& h(\hat{s}(2))\left(\int_0^1 (1-i)(2\Delta-\delta) di - h(1-\hat{s}(2))\int_{\frac{2\Delta-\delta}{3\Delta-\delta}}^1 (1-i)(2\Delta-\delta)\, di\right) \\
&+ h(1-\hat{s}(2))\left(\int_0^1 i(\Delta-\mu\delta) di - h(\hat{s}(2))\int_0^{\frac{2\Delta-\delta}{3\Delta-\delta}} i(\Delta-\mu\delta)\, di\right) \\
=& \tfrac{1}{2}\left(h(1-\hat{s}(2))\left(\Delta-\mu\delta + h(\hat{s}(2))\tfrac{(2\Delta-\delta)((2\Delta-\delta)\mu\delta-\Delta(3\Delta-\delta))}{(3\Delta-\delta)^2}\right) + h(\hat{s}(2))(2\Delta-\delta)\right)
\end{aligned}$$

8.4.4 Ex Post Liability

1. (B harmful) $q_B = 1$ and $\delta_B = \delta$. From Table 1, it can be seen that this generates the same outcome as under a Pigouvian tax.

2. (B harm uncertain) $q_B = 1$ and δ_B remains uncertain with posterior probability of $\tilde{\mu}$: Social welfare is:

$$\begin{aligned}
\mathbb{E}[V^{Liab}(1,1,\delta,\delta)] =& h(1-s)h(s)\left(\int_{\frac{2\Delta-\delta}{4\Delta-(1+\tilde{\mu})\delta}}^1 i(2\Delta-\delta\tilde{\mu}_B)\, di + \int_0^{\frac{2\Delta-\delta}{4\Delta-(1+\tilde{\mu})\delta}} (1-i)(2\Delta-\delta)\, di\right) \\
&+ h(1-s)(1-h(s))\int_0^1 i(2\Delta-\tilde{\mu}\delta)\, di \\
&+ (1-h(1-s))h(s)\left(\int_0^{\frac{2\Delta-\delta}{3\Delta-\delta+\tilde{\mu}\Delta}} (1-i)(2\Delta-\delta)\, di + \int_{\frac{2\Delta-\delta}{3\Delta-\delta+\tilde{\mu}\Delta}}^1 i(\Delta-\tilde{\mu}\delta)\, di\right) \\
&+ (1-h(s))(1-h(1-s))\int_0^1 i(\Delta-\tilde{\mu}\delta)\, di
\end{aligned}$$

3. (B has not advanced) $q_B = 0$ while $\tilde{\mu} = \mu$: Research potentially occurs on both paths in period 2. If neither advances, the social welfare (and private return) will be 0 as A will not be adopted under ex post liability. If A advances while B does not, then $\hat{I}(2) = 1$ and social welfare (and A return) is $\frac{1}{2}(2\Delta - \delta)$. If B advances while A does not, then $\hat{I}(2) = 0$ and the B return is $\frac{1}{2}\Delta$. Finally, if both advance, then $\hat{I}(2) = \frac{2\Delta-\delta}{3\Delta-\delta+\mu\Delta}$ with $v_A(2,1) = \int_0^{\frac{2\Delta-\delta}{3\Delta-\delta+\mu\Delta}} \left((1-i)(2\Delta-\delta) - i\Delta\right) di = \frac{(2\Delta-\delta)^2}{2(3\Delta-\delta+\mu\Delta)}$ and $v_B(2,1) = \int_{\frac{2\Delta-\delta}{3\Delta-\delta+\mu\Delta}}^1 \left(i\Delta - (1-i)(2\Delta-\delta)\right) di = \frac{\Delta^2}{2(3\Delta-\delta+\mu\Delta)}$. At the beginning of period 2, $\hat{s}(2)$ will equate the average returns to scientists researching on each path; that is:

$$\frac{h(\hat{s}(2))}{\hat{s}(2))} \left(h(1-\hat{s}(2)) \int_0^{\frac{2\Delta-\delta}{3\Delta-\delta+\mu\Delta}} \left((1-i)(2\Delta-\delta) - i\Delta\right) di + (1-h(1-\hat{s}(2))) \int_0^1 (1-i)(2\Delta-\delta) di \right)$$

$$= \frac{h(1-\hat{s}(2))}{1-\hat{s}(2))} \left(h(\hat{s}(2)) \int_{\frac{2\Delta-\delta}{3\Delta-\delta+\mu\Delta}}^1 \left(i\Delta - (1-i)(2\Delta-\delta)\right) di + (1-h(\hat{s}(2))) \int_0^1 i\Delta di \right)$$

Thus, at the beginning of period 2, expected social welfare is:

$$\begin{aligned}
\mathbb{E}[V^{Liab}(1,0,\delta,\delta_B)] =& h(\hat{s}(2))h(1-\hat{s}(2)) \left(\int_0^{\frac{2\Delta-\delta}{3\Delta-\delta+\mu\Delta}} (1-i)(2\Delta-\delta)\, di + \int_{\frac{2\Delta-\delta}{3\Delta-\delta+\mu\Delta}}^1 i(\Delta-\mu\delta)\, di \right) \\
&+ h(\hat{s}(2))(1-h(1-\hat{s}(2))) \int_0^1 (1-i)(2\Delta-\delta) di \\
&+ (1-h(\hat{s}(2)))h(1-\hat{s}(2)) \int_0^1 i(\Delta-\mu\delta) di + (1-h(1-s))(1-h(s))0 \\
=& h(\hat{s}(2)) \left(\int_0^1 (1-i)(2\Delta-\delta) di - h(1-\hat{s}(2)) \int_{\frac{2\Delta-\delta}{3\Delta-\delta+\mu\Delta}}^1 (1-i)(2\Delta-\delta)\, di \right) \\
&+ h(1-\hat{s}(2)) \left(\int_0^1 i(\Delta-\mu\delta) di - h(\hat{s}(2)) \int_0^{\frac{2\Delta-\delta}{3\Delta-\delta+\mu\Delta}} i(\Delta-\mu\delta)\, di \right) \\
=& \tfrac{1}{2} h(1-\hat{s}(2)) \left(\Delta - \mu\delta + h(\hat{s}(2)) \tfrac{(2\Delta-\delta)((2\Delta-\delta)\mu\delta - \Delta(3\Delta-\delta+\mu\Delta))}{(3\Delta-\delta+\mu\Delta)^2} \right) \\
&+ \tfrac{1}{2} h(\hat{s}(2))(2\Delta-\delta)
\end{aligned}$$

References

Acemoglu, D. (2011). Diversity and technological progress. In *The Rate and Direction of Inventive Activity Revisited*, pages 319–356. University of Chicago Press.

Acemoglu, D. (2021). Harms of ai. Technical report, National Bureau of Economic Research.

Acemoglu, D. (2023). Distorted innovation: Does the market get the direction of technology right? *AEA Papers and Proceedings*, 113:1–28.

Acemoglu, D. and Johnson, S. (2023). *Power and Progress: Our Thousand-Year Struggle over Technology and Prosperity*. Hachette UK.

Acemoglu, D. and Lensman, T. (2024). Regulating transformative technologies. *American Economic Review: Insights*.

Agrawal, A., Gans, J., and Goldfarb, A. (2022). *Prediction Machines: The Simple Economics of Artificial Intelligence*. Harvard Business Press.

Bryan, K. A. (2017). The perils of path dependence. In Gans, J. S. and Kaplan, S., editors, *Survive and Thrive: Winning Against Strategic Threats to Your Business*. Dog Ear Publishing.

Bryan, K. A. and Lemus, J. (2017). The direction of innovation. *Journal of Economic Theory*, 172:247–272.

Brynjolfsson, E. (2022). The turing trap. *Daedalus*, 151(2):272–287.

Cowan, R. (1990). Nuclear power reactors: a study in technological lock-in. *Journal of Economic History*, 50(3):541–567.

Gans, J. S. (2024). How learning about harms impacts the optimal rate of artificial intelligence adoption. Technical report.

Guerreiro, J., Rebelo, S., and Teles, P. (2023). Regulating artificial intelligence. Technical report, National Bureau of Economic Research.

McLaughlin, C. (1954). The stanley steamer: A study in unsuccessful invention. *Explorations in Entrepreneurial History*, 7.

O'Donoghue, T., Scotchmer, S., and Thisse, J.-F. (1998). Patent breadth, patent life, and the pace of technological progress. *Journal of Economics & Management Strategy*, 7(1):1–32.

Russell, S. (2019). *Human Compatible: AI and the Problem of Control*. Penguin Uk.